Racial and Ethnic Disparities in Occupational Injury Risk and Labor Market Outcomes After Injury: Evidence from California Workers' Compensation

Authors: Michael Dworsky (RAND), Leslie I. Boden (BU SPH), Elizabeth C. Chase (RAND), Yvette Cozier (BU SPH), Sarah Edgington (RAND), Marc Elliott (RAND), and Seth Seabury (USC)

Long Abstract

Nonfatal occupational injuries have significant consequences for injured workers. These include disruption of family life, worse health outcomes such as long-term disability and excess mortality, and economic costs in the form of long-term earnings and employment losses. Although occupational injury rates have declined over time, there were still 2.2 million nonfatal occupational injuries and illnesses involving at least one day lost from work, and past estimates suggest the annual total social cost of these injuries is in the hundreds of billions of dollars. However, relatively little is known about how the costs of these injuries differ according to important individual characteristics, including the race and ethnicity of injured workers.

Racial and ethnic disparities in the burden of nonfatal occupational injury might arise from two sources: disparities in the incidence rate of nonfatal injury, and disparities in the consequences of injury conditional on injury occurrence. Racial and ethnic disparities in the incidence of nonfatal occupational injuries have rarely been measured, however, and there are no estimates of racial and ethnic disparities in the economic consequences of workplace injury (which we define as the causal effect of the injury on earnings and employment). Evidence on these questions is needed to understand how occupational safety contributes to differences in health and economic outcomes across racial and ethnic groups.

This study uses indirect estimation methods to provide new evidence on disparities in occupational health and the economic consequences of injury. We use a unique dataset containing 15 years (2005-2019) of California workers' compensation claims linked to quarterly earnings records covering 2003-2019 from the state unemployment insurance system. Because these administrative records lack direct measures of race or ethnicity, we use a validated algorithm (modified Bayesian Improved First Name-Surname Geocoding, or mBIFSG) to impute race and ethnicity for 10 million workers' compensation claims, allowing us to obtain unbiased estimates of disparities in injury incidence rates and the economic consequences of injury.

We find large disparities in incidence rates for non-fatal injuries resulting in work absence ("lost-time injuries"): Black and Hispanic workers experienced lost-time injury rates that were 74% and 90% higher, respectively, compared to White non-Hispanic workers. Using Poisson regression models with occupation fixed effects to quantify the role of occupational segregation, we found that occupational segregation accounted for 49% of the disparity between Black and White non-Hispanic workers and 69% of the disparity between Hispanic and White non-Hispanic workers. Among Black women, however, most (56%) of the excess risk of injuries was attributed to within-occupation disparities.

To test for disparities in the economic consequences of injury, we will estimate the causal effect of injury on earnings and employment. We will build on a difference-in-differences research design used in past work by comparing workers with lost-time injuries to observably similar workers with minor ("medical-only") injuries that did not result in any compensable work disability. (Identification of causal effects relies on

the assumption that the severity of an injury conditional on occurrence is quasi-random with respect to future labor market outcomes.) We will estimate event-study regression models for earnings and employment from two years before the injury to five years after the injury, stratifying by race, ethnicity, and gender to test for the presence of disparities in earnings. The contribution of observable differences between groups (such as occupation, age at injury and pre-injury earnings) to disparities will be analyzed using Oaxaca-Blinder decomposition methods.

ACKNOLWEDGMENTS

This research was supported by NIOSH award R01OH012578. The authors are grateful to the California Department of Industrial Relations for providing the data used in this study, and to Aaron Kofner and Roald Euller at RAND for assistance with BISG.

Racial and Ethnic Disparities in Occupational Health: Evidence from California Workers' Compensation

Authors: Michael Dworsky*, Ph.D. (RAND), Leslie I. Boden, Ph.D. (BU SPH), Elizabeth C. Chase, Ph.D. (RAND), Yvette Cozier, Ph.D. (BU SPH), Sarah Edgington, M.A. (RAND), Marc Elliott, Ph.D. (RAND), Seth A. Seabury, Ph.D. (USC)
Word Count: 2976/3000
Abstract word count: 281/350
* Corresponding author: mdworsky@rand.org, 310-393-0411, x7141

Racial and Ethnic Disparities in Occupational Health: Evidence from California Workers' Compensation

Key Points

Question: How do occupational injury rates vary with race/ethnicity, and how important is occupational segregation compared to within-occupation disparities in injury risk?

Findings: In this cross-sectional analysis of California workers' compensation claims from 2005-2019, Black and Hispanic workers experienced lost-time injury rates that were 74% and 90% higher, respectively, compared to White non-Hispanic workers. Occupational segregation accounted for 49% of the disparity between Black and White non-Hispanic workers and 69% of the disparity between Hispanic and White non-Hispanic workers.

Meaning: Occupational segregation contributes to occupational health disparities, but within-occupation disparities in risk are also important for public health.

Abstract

Importance: Occupational health disparities significantly impact the safety and well-being of workers across different racial/ethnic groups. Understanding these disparities is crucial for developing targeted interventions, but evidence on occupational injury rates by race/ethnicity has been scarce due to data limitations.

Objective: To determine the incidence of lost-time workplace injuries among major racial/ethnic groups in California and to assess the contributions of occupational segregation and within-occupation disparities to these differences.

Design: Cross-sectional study analyzing of 15 years (2005-2019) of California workers' compensation data. The study examined injury rates across four racial/ethnic

groups: White non-Hispanic, Black non-Hispanic, Asian/Pacific Islander non-Hispanic, and Hispanic.

Setting: California workers' compensation system, encompassing diverse industries and occupations.

Participants: California residents employed in the private sector or state and local government between 2005 and 2019.

Intervention(s) or Exposure(s): Not applicable; the study analyzed existing workers' compensation data without specific interventions.

Main Outcome(s) and Measure(s): The primary outcome was the incidence of lost-time workplace injuries, with comparisons made between racial/ethnic groups. The analysis focused on the role of occupational segregation and within-occupation disparities in explaining injury rates.

Results: Black and Hispanic workers experienced injury rates that were 74% and 90% higher, respectively, compared to White workers, while Asian/Pacific Islander workers had injury rates that were 37% lower. Occupational segregation accounted for 49% of the disparity between Black and White workers and 69% of the disparity between Hispanic and White workers. Notably, 56% of the excess risk of injuries for Black women was attributed to within-occupation disparities.

Conclusions and Relevance: Disparities in workplace safety are a significant contributor to racial/ethnic health disparities. Addressing both occupational segregation and within-occupation disparities is essential for improving workplace safety and reducing health inequities among workers.

Introduction

In 2021-22, employers reported 2.2 million nonfatal occupational injuries and illnesses involving at least one day lost from work. (We use "injuries" to refer to both

injuries and illnesses.)[1] Occupational injuries can lead to disruption of family life,[2,3] lost earnings,[4-6] long-term disability,[7] health problems,[8] and excess mortality.[9,10] The total burden of nonfatal occupational injuries in the United States was estimated to be $198 billion in 2007, or $291 billion in 2023 dollars.[11]

Less is known about how injury risk and the resulting burden vary across racial and ethnic groups. Studies measuring racial/ethnic disparities in nonfatal injury rates or quantifying the role of occupational segregation have reached differing conclusions about the presence and magnitude of racial and ethnic disparities: some found no differences[12], while others found substantial disparities between White workers and Black and Hispanic workers.[13-15] [16]

Occupational segregation into more dangerous jobs could result in higher injury rates for minoritized workers. (We use "occupational segregation" throughout this manuscript to refer to differences in the distribution of occupations between groups without taking a stand on whether these differences are normatively undesirable.) However, differences in job duties, hazard communication, training, or access to protective equipment could result in differences in risk among workers in the same occupation. Although these concerns have been noted for at least 50 years,[17] evidence remains scarce on the magnitude and mechanisms of occupational health disparities. Of the studies that found racial/ethnic disparities in injury rates, two found that adjusting for occupation eliminated disparities,[13,14] while one did not.[18] Studies varied in data sources, case definitions, and the level of detail of occupational classification.[14,16]

Evidence on racial/ethnic differences in workplace injury risk has been hampered by limited data that simultaneously capture exposure (the size of the workforce), injury occurrence, and race/ethnicity. Federal statistics on workplace safety are based on employer-recorded injury logs, but race/ethnicity information is missing in nearly half

of cases,[19] making these data unsuitable for directly measuring racial/ethnic disparities.

We provide new evidence on occupational health disparities using 15 years (2005-2019) of California workers' compensation data. Focusing on nonfatal injuries that result in lost workdays, we examine the incidence of workplace injury across four mutually exclusive major racial/ethnic groups: White, Black, Asian/Pacific Islander (API), and Hispanic (of any race).

Data and Methods

We analyzed workers' compensation administrative claims data and public-use survey data to estimate the rate of nonfatal occupational injuries across racial and ethnic groups, age, sex and occupation. We also quantify the contribution of occupational segregation to differences in risk across groups. The study was approved by the RAND Corporation's Human Subjects Protection Committee (study #2023-N0427).

Data

The data for this study come from two sources: the California Workers Compensation Information System (WCIS) and the American Community Survey (ACS). The WCIS is a database maintained by the California Department of Industrial Relations containing information on all California workers' compensation claims.[20] The ACS is a continuously fielded and nationally representative household survey administered by the US Census Bureau.[21,22]

We used the WCIS to count workplace injuries occurring in 2005-2019 in strata of workers defined by race/ethnicity, gender, age, and occupation (e.g. Black women Fi-

nancial Clerks aged 18-29). We used ACS data from 2005-2019 to estimate the number of full-time equivalent (FTE) workers in the strata described above; details are presented in the Supplement. (One FTE equals 2,000 hours worked.)

Population of Interest and Case Definition

We analyzed the workforce covered by the California workers' compensation system between 2005-2019, focusing on California residents aged 18 to 64. We focused on the rate of nonfatal lost-time injury claims, which we defined to include injured workers who received benefit payments or a settlement for wage-replacement benefits based on temporary or permanent disability.

Missing Data

The WCIS was missing two variables needed for our analysis: the worker's race/ethnicity and the Standard Occupational Classification (SOC) code for the worker's occupation. In addition, the WCIS data had partial missingness on some variables.

We addressed these missing data using multiple imputation .[23] To impute race/ethnicity, we used the modified Bayesian Improved First Name-Surname Geocoding (mBIFSG) algorithm, which uses name and residential address to estimate the probability that people belong to each of six mutually exclusive race/ethnicity categories: White non-Hispanic, Black non-Hispanic, Hispanic, Asian/Pacific Islander non-Hispanic, American Indian/Alaska Native non-Hispanic, and Multiracial non-Hispanic.[24-27] We imputed race/ethnicity for each worker using the probability distribution estimated by mBIFSG. Validation in a general population shows that mBIFSG has very high accuracy for the four racial/ethnic groups we examine in this study.[28]

To obtain SOC codes suitable for use in statistical analysis, we used the NIOSH Industry and Occupation Computerized Coding System (NIOCCS).[29] NIOCCS uses

industry code and occupational text description to probabilistically assign SOC codes. We used NIOCCS to impute occupation for the 95 minor SOC occupations (3-digit SOC codes).

We multiply imputed occupation and race/ethnicity based on multinomial distributions defined by the probabilities estimated by NIOCCS and mBIFSG. We imputed other missing data using Generalized Efficient Regression-Based Imputation with Latent Processes (GERBIL).[30] In each imputed dataset, we counted the number of occupational injuries occurring in each stratification cell grouped by age (ages 18-29, 30-49, 50-64), sex (male/female), race/ethnicity, and occupation. We generated 5 imputations for all estimates, pooling estimates from each imputed dataset and obtaining variance estimates using Rubin's rules. Details are available in the Supplement.

Analysis

Using the number of injuries in each stratum as the outcome and the ACS-based estimate of the number of FTE workers in each stratum as the exposure (offset) term, we used Poisson regression to estimate the association between race/ethnicity and workplace injury risk. We first fit an unadjusted model in which race/ethnicity was the only predictor. We then fit a model that also included gender, age category, and gender interacted with age category. Finally, we estimated a model that added an indicator variable for each occupation.

We used these estimates to calculate adjusted occupational injury risk differences, comparing each minoritized race/ethnicity worker group to White workers. We calculated predicted rates and risk differences using the demographic and occupation distribution observed for White workers. The adjusted risk difference can be interpreted as the risk difference that would result if the demographic and occupation distributions

of, for instance, Black workers were identical to the distributions for White workers. See the Supplement for details

To assess how much of the disparities in injury rates are driven by occupational segregation, we estimated risk differences with and without adjustment for occupation. If adjusting for occupation eliminated the risk differences, we could conclude that disparities are fully attributable to occupational segregation. If adjusting for occupation reduced, but did not eliminate the disparity, we could interpret the percentage change in the risk difference as the proportion of the disparity attributable to occupational segregation. We compare estimated incidence rate ratios across models with and without adjustment for occupation to test whether the disparity attributable to occupational segregation is statistically significantly different from zero.

Finally, we conducted subgroup analyses stratified by sex to examine the possibility that the magnitude of disparities and the role of occupational segregation differed by gender. We also estimated models stratified by occupation, to explore how the racial/ethnic disparity within occupations varied after adjusting for demographic differences.

Results

The data include 2.6 million lost-time injuries over 2005-2019 (**Exhibit 1**). Among injured workers, 37 percent were female and the average age at injury was 42. Age and gender differences between injured workers in different racial/ethnic groups are apparent: women account for a larger share of injured workers among Black and API workers, while the average age at injury for Hispanic workers is nearly 3 years younger than the overall average. Hispanic workers (52.9%) make up more than half

of those injured in California over 2005-2019. The population at risk contains 195 million FTE workers over the 15-year study period .

The population injury rate was 1.3 lost-time injuries per 100 FTE. Lost-time injury rates were far higher for Black (1.74 cases per 100 FTE) and Hispanic workers (1.90 cases per 100 FTE) than for White workers (1.00 cases per 100 FTE), while rates were lower for API workers (0.63 cases per 100 FTE).

Exhibit 2 reports predicted incidence rates and risk differences estimated in our regression models. Full regression tables are in the Supplement.

Adjusting for age and gender results in a slightly larger risk difference between Hispanic workers (2.04-1.00=1.04 vs. 1.90-1.00=0.90, $p < 0.01$) and White workers, but has little impact on the disparities estimated for Black and API workers.

Adjusting for occupation substantially reduces, but does not eliminate, the disparities between Black and Hispanic injury rates and White injury rates. The curly braces in Exhibit 2 report the percentage reduction in the risk difference resulting from adjusting for occupation.

For Black workers, adjusting for occupation reduces the demographic-adjusted risk difference by 53%, from 0.78 injuries per 100 FTE to 0.37 injuries per 100 FTE. For Hispanic workers, adjusting for occupation reduces the demographic-adjusted risk difference by 72%, from 1.04 injuries per 100 FTE to 0.30 injuries per 100 FTE. Changes in the underlying IRR (Incidence Rate Ratio) from adjusting for occupation are statistically significant ($p < 0.01$) for both groups. For API workers, in contrast, the risk difference is unaffected by adjustment for occupation ($p = 0.95$).

To illustrate how occupational segregation results in differential exposure to high-risk occupations, each panel of **Exhibit 3** plots the proportion of workers by race/ethnicity in each of the 95 occupations we analyze against the crude lost-time injury rate

estimated in our data. For White and API workers, the employment share in an occupation declines with the lost-time injury rate. In contrast, the employment share of Black workers is not closely related to the lost-time injury rate. For Hispanic workers, in contrast, the employment share in an occupation rises sharply with the lost-time injury rate. These differences in the relationship between injury risk and employment shares help explain why adjusting for occupation does so much to reduce the risk differences.

Estimated Racial/Ethnic Disparities by Gender

Exhibit 2 also shows adjusted rates stratified by gender. The direction of the unadjusted disparities and the effects of adjustment for age are similar for both male and female workers, with higher injury rates for Black and Hispanic workers and lower injury rates for API workers in comparison to White workers. Similarly, age-adjusted disparities are larger than the unadjusted disparities for Hispanic workers ($p = 0.01$ for men and $p = 0.02$ for women), but not meaningfully different from the unadjusted disparities for Black and API workers.

When we adjust for occupation, however, the results for men and women reveal gender differences in the importance of occupational segregation relative to within-occupation disparities for Black and Hispanic workers.

For men, occupational segregation accounts for most of the demographic-adjusted disparity in both groups. Adjusting for occupation reduces the disparity for Black men by 71%, from 0.62 injuries to 0.18 injuries per 100 FTE, and adjusting for occupation reduces the disparity for Hispanic men by 77%, from 0.99 injuries to 0.23 injuries per 100 FTE.

For women, occupational segregation explains less of the demographic-adjusted disparity. For Hispanic women, occupational segregation still accounts for the majority (63%) of the age-adjusted disparity: adjusting for occupation reduces the disparity from 1.10 injuries to 0.41 injuries per 100 FTE.

For Black women, in contrast, adjusting for occupation explains only 44% of the age-adjusted disparity: adjusting for occupation reduces the disparity from 0.95 cases to 0.53 cases per 100 FTE. Comparing the disparities between men and women within race/ethnicity, the within-occupation risk difference for Black women is nearly triple the within-occupation risk difference for Black men, while the within-occupation risk difference for Hispanic women is 77% higher than the within-occupation risk difference for Hispanic men.

Within-Occupation Disparities

While the results above highlight the importance of occupational segregation as an explanation for the higher injury rates experienced by Black and Hispanic workers in comparison to White workers, the within-occupation disparities for these groups remained substantial--most notably for Black women. We accordingly estimated within-occupation disparities (risk differences adjusted for demographics) for each of the 95 occupations examined in our study by stratifying our regression model on occupation.

Exhibit 4 presents the estimated within-occupation disparities by plotting the injury rate in each occupation adjusted for age and gender (i.e., the incidence rate from the occupation-specific model that would be predicted under the White age and gender distribution) against the rate observed for White workers.

For both Black and Hispanic workers, the estimated within-occupation risk differences are positive for the majority of occupations examined. The demographic-adjusted Black injury rate was statistically significantly greater than the White rate in 61

out of 95 occupations, while the Black injury rate was statistically significantly lower than the White rate in only 5 out of 95 occupations. Likewise, the demographic-adjusted Hispanic injury rate was statistically significantly greater than the White rate in 71 out of 95 occupations, while the Hispanic injury rate was statistically significantly lower than the White rate in only 5 out of 95 occupations.

We conclude that the occupation-adjusted disparities between Black and Hispanic and White workers (**Exhibit 2**) are not driven by a small number of occupations or differentially driven by low- or high-risk occupations. Instead, within-occupation disparities are pervasive, and even some lower-risk occupations have substantial within-occupation disparities that contribute to the overall disparities reported above. Additional detail on the regression estimates, adjusted rates and risk differences for all occupations with at least 1500 FTE and more than 10 injuries for each racial/ethnic group can be found in the Supplement.

Occupation-specific estimates stratified by gender are also presented in the Supplement. These resemble the overall estimates: within-occupation disparities for both men and women overwhelmingly have the same sign as the within-occupation effects reported in Table 2, indicating that within-occupation disparities are widespread and are not driven by a small set of occupations.

Discussion

Our study analyzed how lost-time injury rates varied across racial/ethnic groups in California over 2005-2019, finding that Black and Hispanic workers experienced workplace injuries at far higher rates than White workers. These disparities reflect both occupational segregation and elevated risk compared to workers within the same occupation. We also noted some previously unknown gender differences in the role of

occupational segregation: occupational segregation explains a greater proportion of the overall disparity for men (71% for Black men and 77% for Hispanic men) than for women (44% for Black women and 63% for Hispanic women).

Limitations

One limitation of our study is that many workplace injuries are not recorded by either the SOII or workers' compensation.[31-33] One study using 2007-2008 data from California estimated that workers' compensation captures about 75 percent of all lost workday injuries.[38] However, this is comparable to estimates of the proportion of California reportable injuries captured in the SOII during our study period, suggesting that while our data may be incomplete they likely represent the most comprehensive available set of information on workplace injuries.[42]

We also note that, if anything, this limitation suggests our findings understate the burden of workplace injuries to minoritized populations. If, as suggested by past work, underreporting is associated with dimensions of social vulnerability (like English language proficiency, precarious work arrangements, or the threat of employer retaliation)[39] that are more pronounced in the Black and Hispanic populations, our study may underestimate the true disparities in injury risk between these groups and the White population.

Another limitation of our study is that we did not directly observe race and ethnicity and had to impute it based on other observable factors. However, this limitation is common to most large administrative databases, including healthcare claims, and the methods we employ have been well-validated in prior work. Our methods did not allow us to observe more disaggregated racial/ethnic groups than those included in the

mBIFSG output. In particular, aggregation of the Hispanic and Asian and Pacific Islander populations[34] in this analysis could mask policy-relevant heterogeneity in the occupational exposures and within-occupation injury risks faced by these two groups.

Implications for Policy and Research

We found substantial racial/ethnic disparities in nonfatal injury rates, and we also found that differences in occupational distribution account for much of these disparities. While past research in this area has been inconsistent, our results combined with some other recent studies would suggest these past inconsistencies are likely due to data limitations and/or a focus on small convenience samples. The closest analog to our study, a recent study in Washington State using similar methods to ours, found nonfatal injury rates among Black workers were substantially higher than for Hispanic and White workers for broad occupation and industry categories and also found disparities between Hispanic and White workers for most of these categories.[16] Our work builds on this and other studies by providing a more complete assessment of the extent to which disparities in workplace injury can be explained by occupational segregation.

The fact that risk differences persist even when comparing workers in the same jobs, particularly for Black and female workers, suggests that minoritized workers systematically face worse working conditions and riskier environments. Past work has shown that workplace injuries and exposure to adverse working conditions contribute to poor health and disability at older ages, so our findings suggest that the occupational health disparities identified here could be an important contributing factor to overall health disparities. Efforts to alleviate health disparities should consider the role of working conditions, not just economic outcomes such as income, as a potentially important, modifiable factor.

Our study also has implications for the regulation of workplace safety. Much of the focus by regulatory bodies such as OSHA centers on occupation, as injury risk varies considerably across jobs. But our work suggests that regulators and inspectors should also consider the composition of the workforce and the practices of specific employers to best protect workers. For example, future work could consider whether targeting safety inspections based on workforce demographics could help better identify risky workplaces.

EXHIBIT 1 (table)
CAPTION: Lost-Time Workers' Compensation Claims and FTE Employment in California by Race/Ethnicity and Gender, 2005-2019
SOURCE: Authors' calculations, 2005-2019 WCIS and ACS. NOTES: “API" = Asian/Pacific Islander. "AIAN" = American Indian/Alaska Native. " Multiracial" = 2 or more races. "SD" = Standard Deviation. All groups other than Hispanic are non-Hispanic.

EXHIBIT 2 (figure)
CAPTION Unadjusted and Adjusted Lost-Time Injury Incidence Rates per 100 FTE and Contribution of Occupational Segregation to Demographic-Adjusted Disparity, by Race/Ethnicity and Gender
SOURCE: Authors' calculations, 2005-2019 WCIS and ACS. NOTES: “API" = Asian/Pacific Islander. "FTE" = full-time equivalent. Heteroskedasticity-robust standard errors are in parentheses, and 95% confidence intervals are in brackets. Table reports predicted incidence rates and risk differences from Poisson regression models for lost-time injury counts. The number of FTE workers (in hundreds) is used as an exposure term (i.e., ln(FTE / 100) is included with the coefficient constrained to equal one. All models include an intercept and indicators for the six race/ethnicity categories produced by mBIFSG (estimates for AIANNH and MNH not shown). Demographic-adjusted models also include indicators for gender (excluded category = men), age (excluded category = ages 18-29), and gender-age interactions in models for all workers. Demographic-adjusted models stratified by gender include only age category indicators. Demographic and Occupation-adjusted models also include indicators for 95 occupation categories (excluded category = top executives (SOC 11-1000)). Adjusted incidence rates are predicted injury rates per 100 FTE under the distribution of included co-variates observed for White workers. Risk Differences for each group are differences between the adjusted rate and the rate observed for White workers. The percent changes indicated with curly braces are the percent reduction in the risk difference between the model that adjusts for demographics and occupation and the model that adjusts for demographics only.

EXHIBIT 3 (figure)
CAPTION: Occupational Employment Shares (%) vs. Overall Injury Rates, by Race/Ethnicity
SOURCE: Authors' calculations, 2005-2019 WCIS and ACS. NOTES: Crude injury rate per 100 FTE is calculated for all workers (i.e., averaged over all racial/ethnic groups). Line of best fit reported in figure is a weighted least squares regression of employment shares on the crude injury rate using FTE employment as weights.

EXHIBIT 4 (figure)
CAPTION: Within-Occupation Demographic-Adjusted Injury Rates vs. White Injury Rates, by Race/Ethnicity and Occupation
SOURCE: Authors' calculations, 2005-2019 WCIS and ACS.
NOTES: Dashed line is the 45-degree line. "Significant" (or "Not significant") indicates statistical significance (or in-significance) at the 5% level of the difference between the adjusted rate for the minoritized group and the injury rate among White workers within each occupation.

Exhibit 1: Lost-Time Workers' Compensation Claims and FTE Employment in California by Race/Ethnicity and Gender, 2005-2019

All workers							
	White	Black	Hispanic	API	AIAN	Multiracial	Total
Female (%)	37	49	33	45	41	43	37
Temporary disability benefit received (%)	83	79	83	82	85	83	83
Permanent disability benefit received (%)	46	50	45	45	44	45	46
Indemnity or unspecified settlement received (%)	18	20	26	17	21	18	22
Age at injury	42	42	39	42	42	42	42
Number of lost-time injuries, 2005-2019	794,899	171,256	1,364,180	183,889	9,624	52,590	2,576,438
Proportion of lost-time injuries in column (%)	30.9	6.6	52.9	7.1	0.4	2.0	100
Number of FTE, 2005-2019 (million FTE)	79,529,213	9,849,858	71,700,286	29,284,392	611,326	3,945,926	194,921,001
Proportion of FTE in column, 2005-2019	40.8	5.1	36.8	15	0.3	2	100
Unadjusted Injury Rate per 100 FTE	1.00	1.74	1.90	0.63	1.57	1.33	1.32

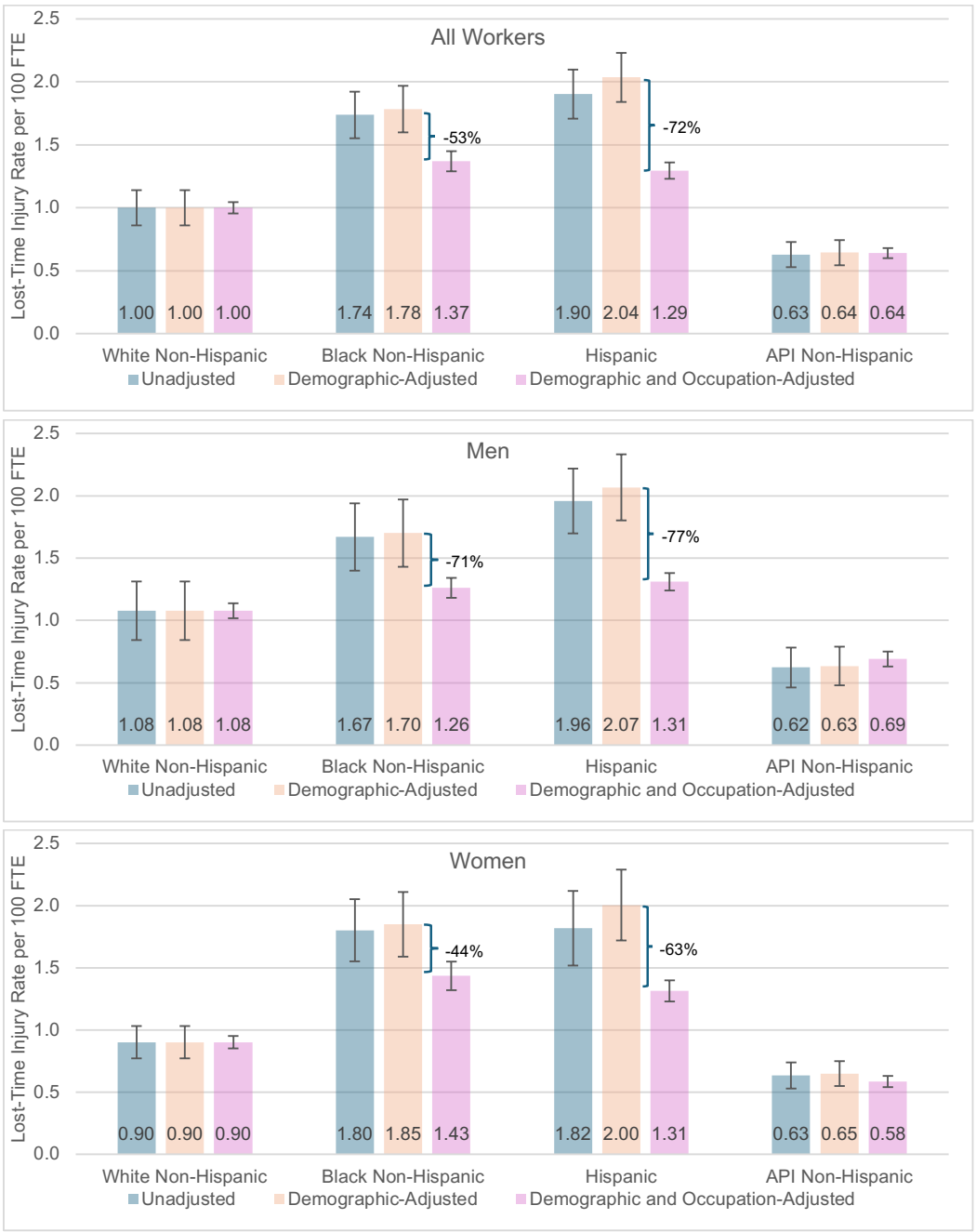
Men							
	White	Black	Hispanic	API	AIAN	Multiracial	Total
Female (%)	0	0	0	0	0	0	0
Temporary disability benefit received (%)	83	80	84	83	86	83	84
Permanent disability benefit received (%)	45	48	43	43	43	43	44
Indemnity or unspecified settlement received (%)	18	20	25	18	22	18	22
Age at injury	42	42	38	41	41	41	41
Number of lost-time injuries, 2005-2019	477,967	80,706	849,115	95,706	5,310	27,750	1,536,554
Proportion of lost-time injuries in column (%)	31.1	5.3	55.3	6.2	0.3	1.8	100
Number of FTE, 2005-2019 (million FTE)	35,258,166	5,025,597	28,395,536	13,934,278	296,623	1,896,003	84,806,203
Proportion of FTE in column, 2005-2019	41.6	5.9	33.5	16.4	0.3	2.2	100
Unadjusted Injury Rate per 100 FTE	1.36	1.61	2.99	0.69	1.79	1.46	1.81

(continued on next page)

Women							
	White	Black	Hispanic	API	AIAN	Multiracial	Total
Female (%)	100	100	100	100	100	100	100
Temporary disability benefit received (%)	83	79	80	81	84	83	81
Permanent disability benefit received (%)	48	53	51	47	45	48	50
Indemnity or unspecified settlement received (%)	18	21	27	17	21	18	22
Age at injury	44	42	41	43	43	43	43
Number of lost-time injuries, 2005-2019	317,769	90,654	516,236	88,333	4,326	24,877	1,042,196
Proportion of lost-time injuries in column (%)	30.5	8.7	49.5	8.5	0.4	2.4	100
Number of FTE, 2005-2019 (million FTE)	44,271,047	4,824,261	43,304,750	15,350,115	314,703	2,049,923	110,114,798
Proportion of FTE in column, 2005-2019	40.2	4.4	39.3	13.9	0.3	1.9	100
Unadjusted Injury Rate per 100 FTE	0.72	1.88	1.19	0.58	1.37	1.21	0.95

NOTES: Authors' calculations, 2005-2019 WCIS and ACS. "API" = Asian/Pacific Islander. "AIAN" = American Indian/Alaska Native. " Multiracial" = 2 or more races. "SD" = Standard Deviation. All groups other than Hispanic are non-Hispanic.

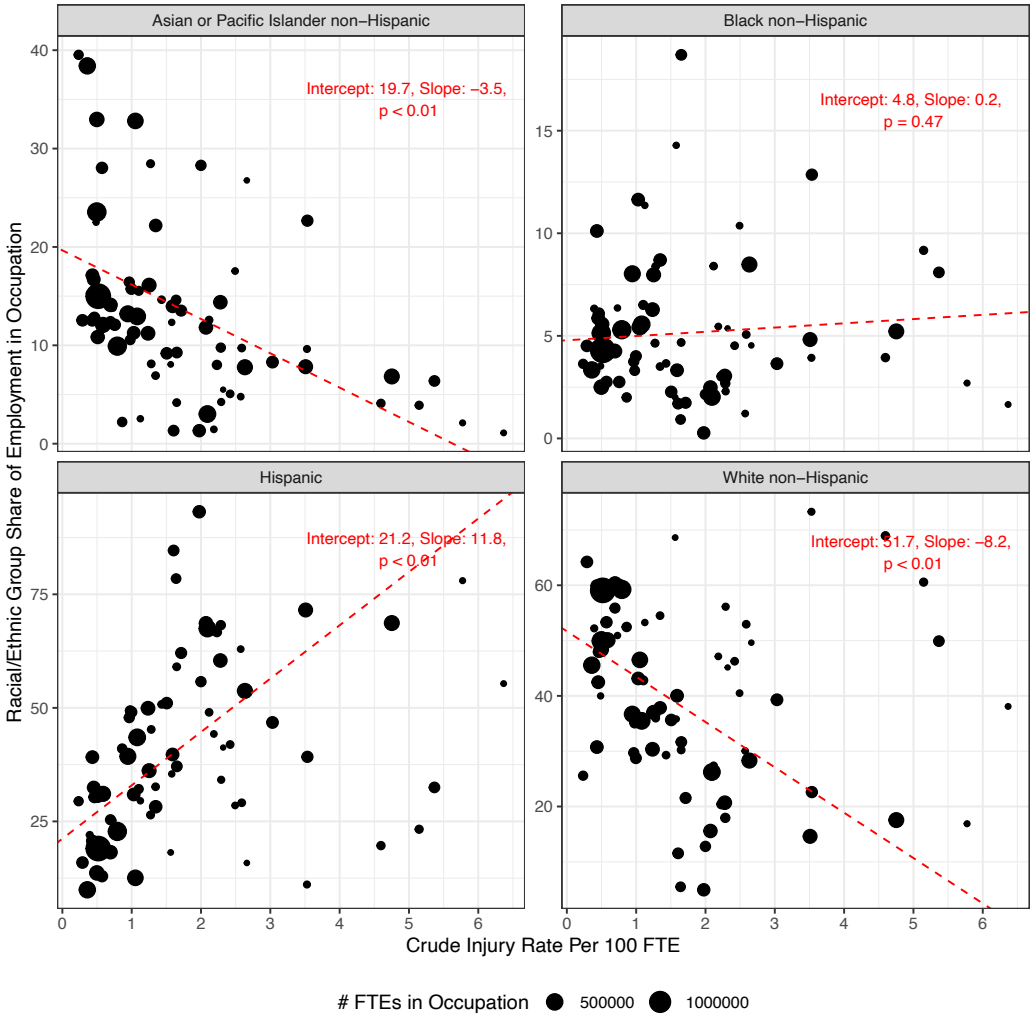
Exhibit 2: Unadjusted and Adjusted Lost-Time Injury Incidence Rates per 100 FTE and Contribution of Occupational Segregation to Demographic-Adjusted Disparity, by Race/Ethnicity and Gender



NOTES: Authors' calculations, 2005-2019 WCIS and ACS. "API" = Asian/Pacific Islander. "FTE" = full-time equivalent. Heteroskedasticity-robust standard errors are in parentheses, and 95% confidence intervals are in brackets. Table reports predicted incidence rates and risk differences from Poisson regression models for lost-time injury counts. The number of FTE workers (in hundreds) is used as an exposure term (i.e., ln(FTE / 100) is included with the coefficient constrained to equal one. All models include an intercept and indicators for the six race/ethnicity categories produced by mBIFSG (estimates for AIANNH and MNH not shown). Demographic-adjusted models also include indicators for gender (excluded category = men), age (excluded category = ages 18-29), and gender-age interactions in models for all workers. Demographic-adjusted models stratified by gender include

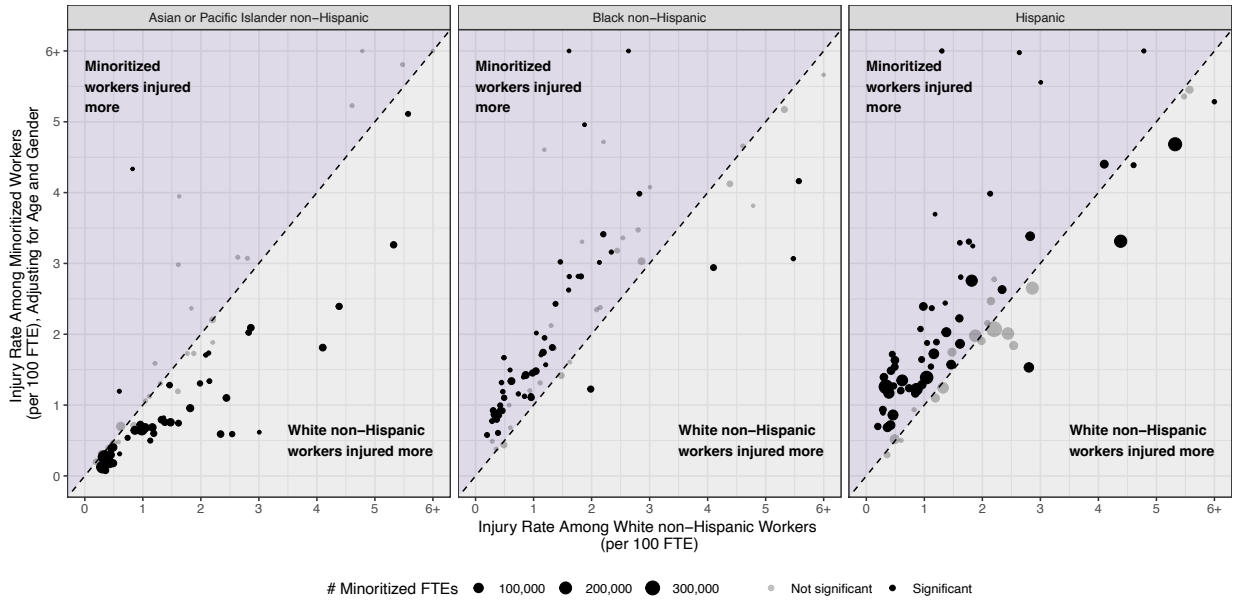
only age category indicators. Demographic and Occupation-adjusted models also include indicators for 95 occupation categories (excluded category = top executives (SOC 11-1000)). Adjusted incidence rates are predicted injury rates per 100 FTE under the distribution of included covariates observed for White workers. Risk Differences for each group are differences between the adjusted rate and the rate observed for White workers. The percent changes indicated with curly braces are the percent reduction in the risk difference between the model that adjusts for demographics and occupation and the model that adjusts for demographics only.

Exhibit 3: Occupational Employment Shares (%) vs. Overall Injury Rates, by Race/Ethnicity



NOTES: Authors' calculations, 2005-2019 WCIS and ACS. Crude injury rate per 100 FTE is calculated for all workers (i.e., averaged over all racial/ethnic groups). Line of best fit reported in figure is a weighted least squares regression of employment shares on the crude injury rate using FTE employment as weights.

Exhibit 4: Within-Occupation Demographic-Adjusted Injury Rates vs. White Injury Rates, by Race/Ethnicity and Occupation



NOTES: Authors' calculations, 2005-2019 WCIS and ACS. Dashed line is the 45-degree line. "Significant" (or "Not significant") indicates statistical significance (or insignificance) at the 5% level of the difference between the adjusted rate for the minoritized group and the injury rate among White workers within each occupation.

References

1. U.S. Department of Labor, Bureau of Labor Statistics. Employer-reported workplace injuries and illnesses - 2021-2022. https://www.bls.gov/news.release/pdf/osh.pdf. Accessed October 18. 2024,
2. Strunin L, Boden LI. Family consequences of chronic back pain. Soc Sci Med. Apr 2004;58(7):1385-1393. doi:Doi 10.1016/S0277-9536(03)00333-2
3. Keogh JP, Nuwayhid I, Gordon JL, Gucer PW. The impact of occupational injury on injured worker and family: outcomes of upper extremity cumulative trauma disorders in Maryland workers. Research Support, U.S. Gov't, P.H.S. Am J Ind Med. Nov 2000;38(5):498-506.
4. Reville RT, Boden LI, Biddle JE, Mardesich C. An evaluation of New Mexico workers' compensation permanent partial disability and return to work. 2001.
5. Seabury SA, Scherer E, O'Leary P, Ozonoff A, Boden L. Using linked federal and state data to study the adequacy of workers' compensation benefits. Am J Ind Med. 2014;57(10):1165-1173.
6. Woock C. Earnings losses of injured men: Reported and unreported injuries. Ind Relat. 2009;48(4):610-628.
7. O'Leary P, Boden LI, Seabury SA, Ozonoff A, Scherer E. Workplace injuries and the take-up of Social Security disability benefits. Social security bulletin. 2012;72(3):1-17.
8. Dong XS, Wang X, Largay JA, Sokas R. Long-term health outcomes of work-related injuries among construction workers—findings from the National Longitudinal Survey of Youth. Am J Ind Med. 2015;58(3):308-318.
9. Boden LI, Asfaw A, Busey A, et al. Increased all-cause mortality following occupational injury: A comparison of two states. OEM. 2022;
10. Martin CJ, Jin C, Bertke SJ, Yiin JH, Pinkerton LE. Increased overall and cause-specific mortality associated with disability among workers' compensation claimants with low back injuries. Am J Ind Med. 2020;63:209-217. doi:10.1002/ajim.23083
11. Leigh JP. Economic burden of occupational injury and illness in the United States. Milbank Q. 2011;89(4):728-772.
12. Oh J-H, Shin EH. Inequalities in nonfatal work injury: The significance of race, human capital, and occupations. Soc Sci Med. 2003;57(11):2173-2182.
13. Strong LL, Zimmerman FJ. Occupational injury and absence from work among African American, Hispanic, and non-Hispanic White workers in the national longitudinal survey of youth. Am J Public Health. 2005;95(7):1226.
14. Leeth JD, Ruser J. Safety segregation: The importance of gender, race, and ethnicity on workplace risk. J Econ Inequal. 2006;4:123-152.
15. Baron SL, Steege AL, Marsh SM, Menendez CC, Myers JR. Nonfatal work-related injuries and illnesses—United States, 2010. CDC Health Disparities and Inequalities Report—United States, 2013. 2013;62(3):35.
16. Smith CK, Wuellner S, Marcum J. Racial and ethnic disparities in workers' compensation claims rates. PLoS One. 2023;18(1):e0280307.
17. Lloyd JW. Long-term mortality study of steelworkers. V. respiratory cancer in coke plant workers. J Occup Environ Med. 1971;13(2):53-68.

18. Marcum J, Adams D. Work-related musculoskeletal disorder surveillance using the Washington state workers' compensation system: Recent declines and patterns by industry, 1999-2013. Am J Ind Med. 2017;
19. U.S. Bureau of Labor Statistics USDoL. Biennial case and demographic characteristics for work-related injuries and illnesses involving days away from work, restricted activity, or job transfer - 2021-22: R71. Selected worker characteristics by number of days by case type (Number). 2024.
20. State of California Department of Industrial Relations. Workers' Compensation Information System. State of California Department of Industrial Relations. Accessed November 5, 2024. https://www.dir.ca.gov/dwc/wcis.htm
21. United States Census Bureau. American Community Survey (ACS). United States Census Bureau. Accessed October 23, 2024. https://www.census.gov/programs-surveys/acs
22. Ruggles S, Flood S, Sobek M, et al. Version 15.0 [dataset]. IPUMS. Accessed November 5, 2024. https://doi.org/10.18128/D010.V15.0
23. Rubin D. Multiple imputation for nonresponse in surveys. John Wiley & Sons, Inc.; 1987.
24. Cabreros I, Agniel D, Martino SC, Damberg CL, Elliott MN. Predicting race and ethnicity to ensure equitable algorithms for health care decision making. Health Aff (Millwood). Aug 2022;41(8):1153-1159. doi:10.1377/hlthaff.2022.00095
25. Elliott M, Becker K, Beckett M, Hambarsoomian K, Pantoja P, Karney B. Using indirect estimates based on name and census tract to improve the efficiency of sampling matched ethnic couples from marriage license data. Public Opin Q. 2013;77(1):375-384.
26. Elliott M, Morrison P, Fremont A, McCaffrey D, Pantoja P, Lurie N. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. Health Serv Outcomes Res Methodol. 2009;9:69-83.
27. Sorbero M, Euller R, Kofner A, Elliott M. Imputation of race and ethnicity in health insurance marketplace enrollment data, 2015-2022 open enrollment periods. 2022.
28. Branham DK, Finegold K, Chen L, et al. Trends in missing race and ethnicity information after imputation in HealthCare.gov marketplace enrollment data, 2015-2021. JAMA Netw Open. 2022;5(6):e2216715-e2216715.
29. NIOSH Industry and Occupation Computerized Coding System (NIOCCS). United States Centers for Disease Control and Prevention, The National Institute for Occupational Safety and Health. Welcome to the NIOSH Industry and Occupation Computerized Coding System (NIOCCS). Accessed October 23, 2024. https://csams.cdc.gov/nioccs/
30. Robbins M. Joint imputation of general data. J Surv Stat Methodol. 2024;12(1):183-210.
31. Rosenman KD, Kalush A, Reilly MJ, Gardiner JC, Reeves M, Luo Z. How much work-related injury and illness is missed by the current national surveillance system? J Occup Environ Med. 2006;48(4):357-365.
32. Boden LI, Ozonoff A. Capture–recapture estimates of nonfatal workplace injuries and illnesses. Ann Epidemiol. 2008;18(6):500-506.
33. Azaroff LS, Levenstein C, Wegman DH. Occupational injury and illness surveillance: Conceptual filters explain underreporting. Am J Public Health. Sep 2002;92(9):1421-1429. doi:Doi 10.2105/Ajph.92.9.1421
34. Shimkhada R, Scheitler AJ, Ponce NA. Capturing racial/ethnic diversity in population-based surveys: Data disaggregation of health data for Asian American,

Native Hawaiian, and Pacific Islanders (AANHPIs). Population Res Pol Rev. 2021;40:81-102.
42. Gunter MM. An update on SOII undercount research activities. Monthly Lab. Rev.. 2016;139:1.

Supplement for Racial and Ethnic Disparities in Occupational Health: Evidence from California Workers' Compensation

Numerator Construction: Workers' Compensation Information System

Our data come from the California Workers' Compensation Information System (WCIS). The WCIS is a database maintained by the Division of Workers' Compensation (DWC) in the California Department of Industrial Relations (DIR) that contains information on all workers' compensation claims.[1] Under a data use agreement, the Division of Workers' Compensation in the California Department of Industrial Relations provided us with WCIS data on all claims with dates of injury from 2005-2019.

California law requires workers compensation claims administrators (insurers, self-insured employers, and third-party administrators) to report all new claims to the WCIS. Like workers' compensation claims databases used in other states, the WCIS is a transaction-based Electronic Data Interchange (EDI) database that generally follows the claims standards defined by the International Association of Industrial Accident Boards and Commissions (IAIABC).

Claim administrators in California are required to submit reports to the WCIS reflecting the occurrence of an injury and subsequent developments, such as benefit payments or case closure. This initial report made upon an injury's occurrence is called the first report on injury (FROI). After the FROI is filed, claims administrators are required to submit additional reports (the subsequent report on injury, or SROI) reflecting material changes in the claim's status, such as claim closure and benefit payments.

Under a data use agreement, programmers at the DWC provided the study team with extracts from FROI and SROI linked to injuries occurring during our 2005-2019 study period. Key data elements from the FROI used directly in the analyses reported here included:

- Gender Code
- Age at Injury (derived from Employee Date of Birth and Date of Injury)

Assignment of Standard Occupation Classification (SOC) codes using the NIOCCS algorithm (discussed below) used two FROI data elements:

[1] More information about the WCIS can be found at https://www.dir.ca.gov/dwc/wcis.htm. Accessed January 23, 2025.

– Occupation Description

– Industry Code (SIC codes were crosswalked to two-digit NAICS codes, and invalid NAICS codes were set to missing before running NIOCCS)

Estimation of race/ethnicity probabilities using mBIFSG used several additional data elements:

– Employee Name (First Name, Middle Name, and Last Name)

– Employee Address (Address, City, State, and Postal code)

Additional data elements from the FROI that were not used in our analysis were also included in our analytic file and used in our multiple imputation process (discussed below). These included:

– Nature, Part of Body, and Cause of Injury Codes

– Date reported to Employer (used to calculate duration between injury date and date claim was reported)

– Date reported to Claim Administrator (used to calculate duration between injury date and date claim was reported)

– Date of Hire (used to calculate job tenure at injury)

– Class Code

– Weekly wage (derived from wage and wage period variables)

– Region within California (a 10 region classification used within the workers' compensation system, which we derived from the injured worker's ZIP code)

– Self-insured status of claim

– Claim denial indicator (constructed by DWC programmers to indicate if claim was denied)

The key SROI data elements used in the current study (including the imputation process) were:

– Paid-to-date amounts indicating the cumulative benefit payments and expenses on the claim at the time of data extraction for a number of detailed benefit type codes

– Date of maximum medical improvement (if any)

Lost-Time Case Definition

Benefit payments reported on subsequent reports of injury are identified using "benefit type codes" (BTCs). We defined lost-time cases to include those with positive paid amounts in any the following BTCs:

- Temporary Disability Benefits
  - 050 Temporary Total
  - 070 Temporary Partial
  - 240 Employer Paid
  - 051 Temporary Total Catastrophic
- Temporary Disability Settlements
  - 550 Compromised Temporary Total
  - 570 Compromised Temporary Partial
  - 524 Compromised Employer Paid
  - 551 Compromised Temporary Total Catastrophic
- Permanent Disability Benefits
  - 020 Permanent Total
  - 030 Permanent Partial Scheduled
  - 090 Permanent Partial Disfigurement
  - 021 Permanent Total Supplemental
  - 040 Permanent Partial Unscheduled
- Permanent Disability Settlements
  - 520 Compromised Permanent Total
  - 530 Compromised Permanent Partial Scheduled
  - 590 Compromised Permanent Partial Disfigurement
  - 521 Compromised Permanent Total Supplemental
  - 540 Compromised Permanent Partial Unscheduled
- Other Settlements
  - 500 Unspecified

Missing Data

SROI records were de-duplicated and paid amounts were aggregated to the groupings of benefit types listed above, so that we obtained a file with one record for each injured worker who had one or more SROI reported. This file was then linked to the FROI using the Jurisdiction Claim Number (JCN), a unique identifier for an injury claim assigned by the WCIS when a FROI is submitted.

After constructing this file, there were several types of missing data which we needed to address prior to being able to calculate the number of injuries in the cells used in our analysis. In particular:

1. Missing race/ethnicity: Race/ethnicity information is not collected on claims; therefore, this information is missing for everyone in the sample. We imputed race/ethnicity using mBIFSG (discussed below).
2. Absence of structured occupation codes: The FROI provides industry codes, but has only a free text description of the worker's occupation that cannot be used directly for statistical analysis. All claimants in the sample are thus missing SOC codes, but have information about their occupation contained in the occupation description field. We imputed SOC codes using the NIOCCS auto-coder (discussed below).
3. Missingness of benefit payments due to low-quality SROI reporters: Compliance with reporting requirements among the nearly 500 claim administrators in the California workers' compensation system is high but imperfect. While reporting of FROI (which record occurrence of an injury and contain most of the variables needed for our analysis) is essentially complete,[2] some claim administrators rarely or never submit SROI. Because the SROI are needed to capture benefit payments, it is necessary to impute benefit payment information needed to identify lost-time claims for cases submitted by the low-quality SROI reporters.
4. Sporadic missingness: The WCIS data exhibit occasional missingness due to incomplete data entry on claim forms, e.g. claimant's gender is not reported, or the industry code provided is invalid. This is the "standard" sort of missing data that we expect to arise in any kind of dataset. We expect this missingness to be missing at random, and it is comparatively minor (an overview of the percent missing by variable and year is provided below).

These four missing data challenges were addressed via multiple imputation, using three algorithms. Modified Bayesian Improved First Name-Surname Geocoding (mBIFSG) was used to impute race/ethnicity, NIOSH Industry and Occupation Computerized Coding System

[2] California Department of Industrial Relations. (2018). Workers' compensation information system (WCIS): Aggregate FROI and SROI Data.

(NIOCCS) was used to assign occupation codes for use in our analysis, and Generalized Efficient Regression-Based Imputation with Latent Processes (GERBIL) was used to address both missing benefit payment information due to low-quality SROI reporters and other sporadic missingness.

Imputation of Race/Ethnicity: mBIFSG

To impute race/ethnicity, we use the mBIFSG algorithm, which uses the information in an individual's name and home address (geocoded to the census block group level) to estimate the probability that the individual's self-reported race/ethnicity would be classified in one of six mutually exclusive categories: Black non-Hispanic, White non-Hispanic, Hispanic, Asian/Pacific Islander non-Hispanic, American Indian/Alaska Native non-Hispanic, or Multiple Race non-Hispanic.[3]

Even after running our analytic data through mBIFSG, there are still some individuals with missing race/ethnicity probability vectors because their name and address were missing from the WCIS. This occurred in only 0.4% of cases. These individuals have race/ethnicity probabilities imputed via GERBIL when we address the sporadic missingness, discussed in the next section. Our approach to converting the mBIFSG probability vectors into a multiply imputed race/ethnicity is discussed below after we describe our approach to sporadic missingness and low-quality SROI reporters.

Validity of Estimates Based on mBIFSG

mBIFSG has very high accuracy for the four largest racial/ethnic groups in the US, where accuracy is defined in terms of the algorithm's ability to predict which race and ethnicity an individual would self-report if asked: a recent comparison of predictions to actual self-reports for

[3] For more information on mBIFSG and its predecessor algorithm BISG, please see:

Sorbero, ME, R Euller, A Kofner, and MN Elliott. *Imputation of Race and Ethnicity in Health Insurance Marketplace Enrollment Data, 2015-2022 Open Enrollment Periods.* RAND Corporation (Santa Monica, CA: 2022).

Cabreros, I., D. Agniel, S. C. Martino, C. L. Damberg, and M. N. Elliott. "Predicting Race and Ethnicity to Ensure Equitable Algorithms for Health Care Decision Making." *Health Aff (Millwood)* 41, no. 8 (Aug 2022): 1153-59. https://doi.org/10.1377/hlthaff.2022.00095. https://www.ncbi.nlm.nih.gov/pubmed/35914194.

Elliott, Marc N., Kirsten Becker, Megan K. Beckett, Katrin Hambarsoomian, Philip Pantoja, and Benjamin Karney. "Using Indirect Estimates Based on Name and Census Tract to Improve the Efficiency of Sampling Matched Ethnic Couples from Marriage License Data." *Public Opinion Quarterly* 77, no. 1 (2013): 375-84. https://doi.org/10.1093/poq/nft007.

Elliott, Marc N., Peter A. Morrison, Allen Fremont, Daniel F. McCaffrey, Philip Pantoja, and Nicole Lurie. "Using the Census Bureau's Surname List to Improve Estimates of Race/ Ethnicity and Associated Disparities." *Health Services and Outcomes Research Methodology* 9, no. 2 (2009): 69-83. https://doi.org/10.1007/s10742-009-0047-1.

mBIFSG found C-statistics of 0.96 for Hispanic, 0.96 for Asian/Pacific Islander, 0.95 for Black, and 0.94 for White.[4]

The key assumption necessary for mBIFSG to deliver accurate estimates is that conditional on an individual's true (self-reported) race and ethnicity, their name (first name and surname) and the racial or ethnic composition of their block group are statistically independent. Under this assumption, mBIFSG optimally combines (using Bayes' rule) the information about race and ethnicity contained in geographic, first name, and surname information.

A further assumption necessary for unbiased estimation in our setting is that, conditional on an individual's name and block group, the probability of workers' compensation claim filing is independent of the individual's true (self-reported) race/ethnicity. We were unable to evaluate this assumption in our sample, but validation results indicating good performance for the four largest racial/ethnic groups were found for workers' compensation claims in Washington State.[30,31] Our use of the newer mBIFSG algorithm should improve on the accuracy found in that validation study.

Imputation of Occupation (SOC) Codes: NIOCCS

To assign SOC codes based on the free-text occupation descriptions contained in the WCIS, we use the NIOCCS autocoder, an algorithm developed by NIOSH to improve the efficiency and reproducibility of occupation coding. NIOCCS is a neural net trained on a large corpus of occupation and industry descriptions taken from sources including death certificates and responses to the Behavioral Risk Factor Surveillance System (BRFSS) survey. If an individual only has occupation description (but is missing industry), NIOCCS will still generate a vector of SOC code probabilities.

The version of NIOCCS shared with the study team for use in our secure computing environment takes industry codes and free-text occupation descriptions as inputs and returns a vector of probabilities for each individual, which gives a probability distribution over 808 detailed (6-digit) SOC codes. (Some detailed SOC codes are aggregated to a higher level.)

In addition to generating the probabilities that a trained coder would assign a record to each possible SOC code, NIOCCS can also assign some probability mass to an “Insufficient Information" category, indicating that there is not enough information to assign this person to a particular occupation. To make things computationally feasible, we restrict the probability vector to be

[4] Branham, D Keith, Kenneth Finegold, Lucy Chen, Melony Sorbero, Roald Euller, Marc N Elliott, and Benjamin D Sommers. "Trends in Missing Race and Ethnicity Information after Imputation in Healthcare.Gov Marketplace Enrollment Data, 2015-2021." *JAMA Netw Open* 5, no. 6 (2022): e2216715-e15.

length 30, corresponding to the top 30 most likely occupations (potentially including "insufficient information"). For the vast majority (95%) of the sample, the top 30 occupations comprised 90% or more of their probability mass.[5]

Even after running our analytic file through NIOCCS, we still have some individuals who do not have an occupation probability vector, because they were missing the occupation text description. (This affects 0.4% of the analytic sample.) For individuals who had more than 75% of their probability mass in "Insufficient Information"—an issue that affected 10.1% of cases in the analytic sample—we set their probability vector to missing and imputed occupation probabilities at the 2-digit (Major Occupation) level using GERBIL. Our approach to converting NIOCCS probabilities into a multiply imputed occupation (SOC code) is discussed below after we describe our approach to sporadic missingness and low-quality SROI reporters.

Low-Quality SROI Reporters and Sporadic Missingness: GERBIL

We used GERBIL to address the remaining two missing data issues described above: sporadic missingness and missing benefit payment information due to low-quality SROI reporters. We describe GERBIL after providing additional background on the nature and scope of these two missing data problems.

Low-Quality SROI Reporters

Although California law requires claims administrators to file both the first report of injury (FROI) and subsequent report of injury (SROI), a nontrivial number of claims administrators regularly file only the FROI and not the SROI, possibly due to a lack of training, resources, or interest in complying with the state's reporting requirements. The SROI is critical for providing information on the outcome of the claim that is unknown at the time when the FROI is submitted. Most important for us, the SROI is the source of all information about benefit payments, which we use to identify lost-time claims. Failure to address incomplete SROI reporting will thus result in misclassification of lost-time claims as medical-only claims.

We built on methods developed in past work[6] using WCIS data to identify low-quality SROI reporters by calculating the proportion of claims submitted by each claim administrator that had paid or settled indemnity benefits (which requires submission of a SROI). Claims administrators who reported 100 or fewer FROI over our study period were omitted from this analysis and treated as low-quality SROI reporters, but these claim administrators account for only a trivial number of FROI. After examining the distribution of the indemnity benefit receipt rate, we classified claims administrators for which less than 15% of FROI resulted in indemnity payments as low-quality SROI reporters. (In our past work, we found that cutoffs of 10% or 20% indemnity

[5] For more information on how NIOCCS works, please see https://csams.cdc.gov/nioccs/About.aspx.

[6] Dworsky, Michael, Stephanie Rennane, and Nicholas Broten, Wage Loss Monitoring for Injured Workers in California's Workers' Compensation System: 2013 Injury Year Findings (First Interim Report). Santa Monica, CA: RAND Corporation, 2018. https://www.rand.org/pubs/research_reports/RR2572.html. Also available in print form.

claims yielded very similar results, as did a requirement based on the percent of cases with any SROI reported.) For context, 60% of FROI in the workers' compensation system with injury dates between 2010 and 2023 had one or more SROI reported, of which 23% had one or more SROI indicating indemnity payments reported.

To prepare the data for imputation of benefit payment information, receipt of benefits (used to ascertain lost-time cases) was coded as missing for all claims from the low-quality SROI reporters. We have some concern that benefit payment information that is missing due to low-quality SROI reporters is not missing completely at random, i.e., that the probability of missingness and the true values of the missing variables may be correlated with some variables used in our analysis. We are concerned about this because large commercial insurers and the quasi-public State Fund, which account for most claims in the system, have high data quality, while smaller self-insured and self-administered plans, including some local government employers, are more likely to be low-quality SROI reporters. However, given the richness of the information contained on the FROI (as described above), we think it is reasonable to assume that benefit payment information that is missing due to low-quality SROI reporters is missing at random after conditioning on the other variables in our dataset (including age, gender, occupation, employer's industry, type of injury, and sub-state region, among other variables).

Overview of GERBIL

We use GERBIL, a sophisticated approach for jointly imputing missing-at-random data, to produce multiply imputed datasets that can be used for our analysis. GERBIL was developed to address theoretical and computational limitations of the more well-known Multiple Imputation with Chained Equations (MICE) approach. Key advantages of GERBIL include the ability to accommodate arbitrary variable types (including binary, categorical, and other variables with restricted ranges) without requiring variables to be normally distributed, as well as formulation of the data generating process that guarantees existence of a joint probability distribution over the included variables. MICE, in contrast, imposes stronger parametric restrictions on the included variables but can yield results that are inconsistent with any possible joint distribution, which can lead to instability in the resulting imputation datasets. GERBIL specifies all variables as functions of jointly normal latent variables, using suitable transformations (such as copulas for continuous variables) where needed. Details on these transformations are given in Robbins et al. (2013) and Robbins (2014).[7]

[7] Robbins, MW, SK Ghosh, and JD Habiger. "Imputation in High-Dimensional Economic Data as Applied to the Agricultural Resource Management Survey." *Journal of the American Statistical Association* 108, no. 501 (2013): 81-95.

Robbins, MW. "The Utility of Nonparametric Transformations for Imputation of Survey Data." *Journal of Official Statistics* 30, no. 4 (2014): 675-700.

After transformation, GERBIL assumes all variables are latently joint multivariate normal, takes imputation draws from this joint distribution, and then back-transforms to return all imputed values to the observed scale. Robbins (2024) is the authoritative source on GERBIL and should be consulted for technical details.[8]

GERBIL Implementation Details

We feed the FROI and SROI variables listed above into GERBIL along with the vectors of probabilities returned by NIOCCS and mBIFSG. Most of these variables are not used in our analysis, but are included because they are predictive of variables that are used in our analysis. For example, the nature, cause, and body part of injury (which are reported in essentially all cases) are important predictors of whether the injury will result in disability benefit payments (which is missing on cases reported by low-quality SROI reporters); likewise, the industry, region of California, age, and gender of the worker are likely predictive of race/ethnicity and occupation in the small number of cases where mBIFSG and NIOCCS did not return usable probabilities.

Table A.1 describes the proportion of cases (unique FROI) that were missing data on the key variables used in our analysis. Missingness rates are very low–below 1%–for variables other than the NIOCCS probabilities and benefit amounts (which we set to missing for the low-quality SROI reporters).

Table A.1: Rates of Missing Data Before Multiple Imputation

Variable	% of observations with missing data
Gender Code	0.03
Age at Injury	0.98
Industry (2-digit NAICS)	0.98
Occupation Description	0.41
NIOCCS Result Probabilities	10.5
mBIFSG Result Probabilities	0.03
Benefit receipt and paid amounts (low-quality SROI reporters)	14.9

The variables that are fed into the algorithm (and multiply imputed by GERBIL when missing) are:

– Binary variables: Worker gender; self-insured status; whether a maximum medical improvement date was reported; and whether the claim was denied

[8] Robbins, MW. "Joint Imputation of General Data." *J Surv Stat Methodol* 12, no. 1 (2024): 183-210.

– Categorical variables: industry (2-digit NAICS code); nature of injury; cause of injury; body part of injury; region of California where employee works; calendar quarter of the injury

– Continuous variables: age at injury, workers compensation insurance premium, pre-injury weekly wage, race/ethnicity probabilities produced by mBIFSG, occupation probabilities produced by NIOCCS

– Semicontinuous variables: duration from injury to claims administrator report; duration from injury to employer report; duration from injury to date of maximum medical improvement; and paid benefit amounts in 9 mutually exclusive categories

Semicontinuous variables are variables that are mostly continuous, but that may have a large point mass at a particular value. For example, the amount of permanent disability benefit received is $0 for individuals who never receive permanent disability—the majority of claims–while for individuals who do receive permanent disability benefits, the benefit amount is positive and highly skewed. Similarly, the duration from injury to report of injury are usually semicontinuous, with large point masses at 0, but some claims have a longer lag between injury and report. Finally, there are variables like duration from injury to maximum medical improvement (MMI), which are only observed for individuals who achieve MMI; for individuals who do not, the duration variable is not defined. We accommodate the MMI date in GERBIL by coding individuals who have not achieved MMI as having -99 for this duration value.

Some other variables listed above need further clarification as well.

Workers' compensation premium: We use the workers' compensation pure premium rate for 2020 proposed by the state's workers' compensation actuaries (the Workers' Compensation Insurance Rating Bureau) as a way to parametrize information about the riskiness of occupations (and thus the likely severity of injuries conditional on observed variables) contained in the workers' compensation class code.[9] Workers' compensation class codes are the basic risk classifications used in workers' compensation pricing, and are a required data element on the FROI for fully insured workers' compensation claims. They are also reported voluntarily on many, but not all, self-insured claims.

Benefit Amounts: We aggregate benefit type codes reflecting disability benefits and settlements into five mutually exclusive categories: temporary disability benefits, permanent disability benefits, death benefits, indemnity settlements, and medical settlements. We also include four additional benefit types that are not used in our analysis, but that capture medical costs and thus may contain additional information about injury severity or other worker and injury characteristics: hospital payments, physician payments, other provider payments, and prescription drug payments.

[9] Source: https://www.wcirb.com/sites/default/files/documents/jan2020_approved_pprs_website.xls. Rates for earlier years are not currently available on the WCIRB website, but the relative riskiness of different classifications is unlikely to change dramatically over time.

Occupation Probabilities: Due to computational constraints, we impute the probability of falling into a given 2-digit SOC code (Major Occupation Group) rather than the 3-digit SOC codes (Minor Occupations) used in our analysis. This shortened the occupation probability vector from 76 to 24 elements, which was computationally manageable. Therefore, for individuals who had a missing occupation probability vector, we can only impute their 2-digit SOC code. After GERBIL, but prior to multiple imputation of occupation, we allocated the probability mass in each Major Occupation Group to its constituent Minor Occupations based on the proportions observed among complete records cases (i.e., the 90% of cases that had usable NIOCCS results).

We generated M = 5 imputation datasets from GERBIL, allowing GERBIL's Markov chain Monte-Carlo (MCMC) sampler to run for 150 iterations. Based on our review of trace plots, comparison of imputed and observed mean values, and other MCMC diagnostics, we judged that 150 iterations was enough for the chains to converge. Imputation was conducted separately in each of our 15 years of claims data, building additional flexibility into the imputation process.

Multiple Imputation of Race/Ethnicity and Occupation Codes

After GERBIL, every record in each of our 5 imputation datasets has complete probability vectors for race/ethnicity and occupation. We used these probability vectors to draw race/ethnicity and occupation codes. Because the probabilities generated by GERBIL were constrained to lie between 0 and 1, but we did not impose a constraint requiring them to sum to 1 across categories, we had to renormalize the GERBIL results for race/ethnicity and NIOCCS probabilities. For race/ethnicity, we renormalized the probabilities to sum to 1 for all individuals. For occupation probabilities, we first dropped "Insufficient Information" as a category (removing its probability mass) and then renormalized the probabilities for the remaining occupations to sum to 1.

After renormalizing the occupation and race/ethnicity probability vectors on each imputation, we assign the SOC code and race/ethnicity in each imputation by drawing from a multinomial distribution with probabilities given by the probability vectors. These draws, which were independent of each other and of the GERBIL imputations, serve as each individual's imputed occupation and race/ethnicity in each imputation dataset.

After these steps, data are complete on each imputation and the information we need to construct numerators (how many injuries occurred in a particular category of worker) is ready.

Numerator Aggregation

The steps above give us imputation datasets in which the unit of observation is an individual workers' compensation claim. We need to aggregate this microdata to units of observation corresponding to the denominator estimates (discussed below) that we calculate from the ACS. As described in the paper, we count the number of lost-time injuries over 2005-2019 in cells defined by:

– Age at injury (3 categories): 18-29, 30-49, 50-64

– Gender (2 categories): male, female

– Race/ethnicity (6 categories): Black non-Hispanic, White non-Hispanic, Hispanic, Asian/Pacific Islander non-Hispanic, American Indian/Alaska Native non-Hispanic, or Multiple Race non-Hispanic

– Occupation: 95 categories: 3-digit (Minor Occupation) SOC code.

For sensitivity analyses, we also use alternative case definitions and aggregate them to the cells defined above. These are discussed below when we present sensitivity analysis results.

Denominator Construction: American Community Survey

In addition to numerators (how many injuries occurred in a particular category of worker), we also need denominators: how many workers in a particular category were at risk of injury?

To generate these denominators, we use data from the American Community Survey (ACS). We use ACS data from 2005-2019 (corresponding to our years of claims data) from individuals who live in California. We then further restrict the sample to individuals who work in California, are neither self-employed nor federal employees, who are ages 18-64, are not "Other race" (since none of our numerators consider other race, which is not one of the BISG race/ethnicity categories), and do not report a military occupation. After making sample exclusions, we had a sample of 1,913,064 ACS respondents between 2005-2019. The impact of these exclusions is shown in Table A.2:

Table A.2: ACS Sample Construction

Exclusion	Number of ACS Respondents	Weighted Population Size
Starting Sample: All Residents of CA	4,150,173	427,649,870
Restricting to Those Who Work in CA	2,379,822	253,107,729
Removing Self-Employed (Not Incorporated) Workers	2,176,671	232,193,748
Removing Self-Employed (Incorporated) Workers	2,089,918	223,805,452
Removing Federal Government Employees	2,020,788	216,946,802
Removing Unpaid Family Workers	2,016,110	216,499,456
Restricting to Ages 18-64	1,917,463	207,866,249
Removing Other Race Individuals	1,913,072	207,320,589
Removing Individuals from Military Occupations	1,913,064	207,319,945
Final ACS Sample Size	1,913,064	207,319,945

ACS Estimation

As with the numerators, we need to count the number of full-time equivalent workers in each stratified cell based on age, sex, race/ethnicity, and occupation. First, we calculate the annual hours worked for each ACS respondent by multiplying their report on their "usual hours worked in a week" by the number of weeks they worked in the last year. Note that the number of weeks worked in the last year was intervalled (not continuous) for the ACS in 2008-2018; for respondents from these years, we use the midpoint of the interval as the estimate for how many weeks they worked in the last year. After calculating the annual hours worked for each respondent, we calculate the total annual hours worked in each stratification cell, adjusting for the ACS weights. We divide the total annual hours worked in each stratification cell by 2000 to obtain full-time equivalents.

Table A.3 (next page) reports the age and gender breakdown by race/ethnicity in the California workers' compensation-covered workforce over 2005-2019.

Table A.3: Denominator Summary Statistics and Marginal Probabilities

All workers														
	White		Black		Hispanic		API		AIAN		Multirace		Total	
	N FTE	%	N FTE	%	N FTE	%	N FTE	%	N FTE	%	N FTE	%	N FTE	%
Sex														
Female	2,350,544	44	335,040	51	1,893,036	40	928,952	48	19,775	49	126,400	48	5,653,747	44
Male	2,951,403	56	321,617	49	2,886,983	60	1,023,341	52	20,980	51	136,662	52	7,340,987	56
Age group														
18-29	1,028,386	19	140,139	21	1,400,002	29	359,715	18	8,528	21	81,107	31	3,017,877	23
30-49	2,546,309	48	336,799	51	2,514,264	53	1,089,890	56	19,845	49	133,396	51	6,640,504	51
50-64	1,727,252	33	179,719	27	865,753	18	502,688	26	12,381	30	48,559	18	3,336,352	26
Men														
	White		Black		Hispanic		API		AIAN		Multirace		Total	
	N FTE	%	N FTE	%	N FTE	%	N FTE	%	N FTE	%	N FTE	%	N FTE	%
Age group														
18-29	548,920	19	69,954	22	830,785	29	186,099	18	4,472	21	40,489	30	1,680,718	23
30-49	1,471,382	50	166,029	52	1,546,263	54	584,078	57	10,423	50	71,346	52	3,849,521	52
50-64	931,101	32	85,634	27	509,936	18	253,164	25	6,085	29	24,827	18	1,810,747	25
Women														
	White		Black		Hispanic		API		AIAN		Multirace		Total	
	N FTE	%	N FTE	%	N FTE	%	N FTE	%	N FTE	%	N FTE	%	N FTE	%
Age group														
18-29	479,466	20	70,185	21	569,217	30	173,616	19	4,057	21	40,618	32	1,337,159	24
30-49	1,074,927	46	170,770	51	968,001	51	505,812	54	9,422	48	62,050	49	2,790,983	49
50-64	796,151	34	94,084	28	355,818	19	249,523	27	6,296	32	23,732	19	1,525,605	27

Source: authors' calculations, 2005-2019 ACS

Estimation Details

To assess the effect of race/ethnicity on the risk of occupational injury, we fit a series of Poisson regressions. The first model, which measures the crude association between race/ethnicity and occupational injury risk, was of the form:

$$\log E(injuries_c \mid race/ethnicity_c) = \beta_0 + \beta_1 race/ethnicity_c + \log \text{FTE}_c \quad (1)$$

i.e. a Poisson regression using the number of injuries in a given stratification cell c as the outcome, conditional on the race/ethnicity of the workers in cell c and the number of full-time equivalents (FTE) in cell c as the offset. Note that $\exp \beta_1$ gives the occupational injury risk ratio between workers of a given race/ethnicity and non-Hispanic White workers (the reference category).

To further disentangle the effect of race/ethnicity from other confounders, we then fit a demographic-adjusted version of Model (1):

$$\log E(injuries_c \mid race/ethnicity_c, sex_c, age_c) = \beta_0 + \beta_1 race/ethnicity_c + \beta_2 sex_c + \beta_3 age_c + \log \text{FTE}_c \quad (2)$$

Finally, to assess the effect of occupation, we fit a model that adjusted for it:

$$\log E(injuries_c \mid race/ethnicity_c, sex_c, age_c, occupation_c) = \beta_0 + \beta_1 race/ethnicity_c + \beta_2 sex_c + \beta_3 age_c + \beta_4 occupation_c + \log \text{FTE}_c \quad (3)$$

We used a pseudo-maximum likelihood estimator with heteroskedasticity-robust standard errors, which allows consistent estimation and valid inference in the presence of overdispersion and other sources of arbitrary heteroskedasticity as long as the conditional expectation of the outcome given the explanatory variables is correctly specified.

After fitting all models and calculating predicted rates and risk differences adjusted for included covariates, estimates from the 5 imputed datasets were pooled using Rubin's Rules (Rubin 1987). Additional details on the analysis can be found in the Appendix. Analyses were carried out using Stata version 18.0[32] and R version 4.4.1[33].

Post-Estimation: Adjusted Rates, Risk Differences, Inference, and Hypothesis Testing

Using the results of each of these models, we calculated an expected number of occupational injuries per 100 FTEs as if all workers were non-Hispanic White. That is, we calculated the average predicted incidence rate in the White non-Hispanic subsample, and then manipulated the race/ethnicity variable to obtain (counterfactual) adjusted rates for the other racial/ethnic groups. Risk differences were calculated as the difference between each racial/ethnic group's adjusted rate and the White non-Hispanic rate.

Variance estimation for the Poisson regressions is done using heteroskedasticity-robust standard errors. Standard errors for post-estimation quantities like incidence rate ratios, adjusted incidence rates, and risk differences are calculated using the delta method.

In the paper, we interpret the change in the risk difference between the age/gender-adjusted model and the occupation-adjusted model as the component of the age/gender-adjusted disparity that is attributable to occupational segregation. We test the hypothesis that occupational segregation contributes to the disparity in injury risk by conducting a Wald test for equality of the occupation-adjusted and age/gender-adjusted coefficients (which correspond to the adjusted difference in each racial/ethnic group's injury rate from the White non-Hispanic rate). Hypothesis testing across equations was done using Stata's suest command (which effectively stacks the estimating equations from each model into a seemingly unrelated regression objective function) with standard errors clustered at the cell level to account for covariance of the estimates across equations.

Subgroup Analyses and Stratification

Because we saw evidence of intersectional associations between race and gender, we also fit gender-stratified variants of models (1)-(3). We then calculated the crude and adjusted disparities between race/ethnicity groups separately for men and women.

Finally, to further explore the effect of occupation, we fit occupation-stratified variants of Models (1) and (2) and calculated the crude and demographic-adjusted racial/ethnic disparity in occupational injury risk within each occupation. We also test equality of male and female occupation-adjusted disparities by testing equality of the regression coefficients across models as described above.

Accounting for Multiple Imputation

All analyses described above were conducted within a framework of multiple imputation. On each of our m=5 imputations, we carried out the analysis. We then pooled the results from each of the m=5 imputations according to Rubin's Rules, i.e. took the mean across the 5 imputations as our point estimate and calculated the variance in a way that reflected the additional uncertainty introduced by the imputation procedure.[10] For all estimates/statistical quantities described above, pooling was the last step of the estimation process.

Supplementary Results

Additional Figures Documenting Occupational Segregation

Figures A.1, A.2, and A.3 present the racial/ethnic composition of occupations in California over 2005-2019 including the two smaller racial/ethnic groups omitted from Exhibit 3 in the paper. Occupations are shown in descending order of the White non-Hispanic employment share.

[10] Rubin, DB. *Multiple Imputation for Nonresponse in Surveys.* New York, NY: John Wiley & Sons, Inc., 1987.

Figure A.1: Racial/Ethnic Composition by Occupation: Total

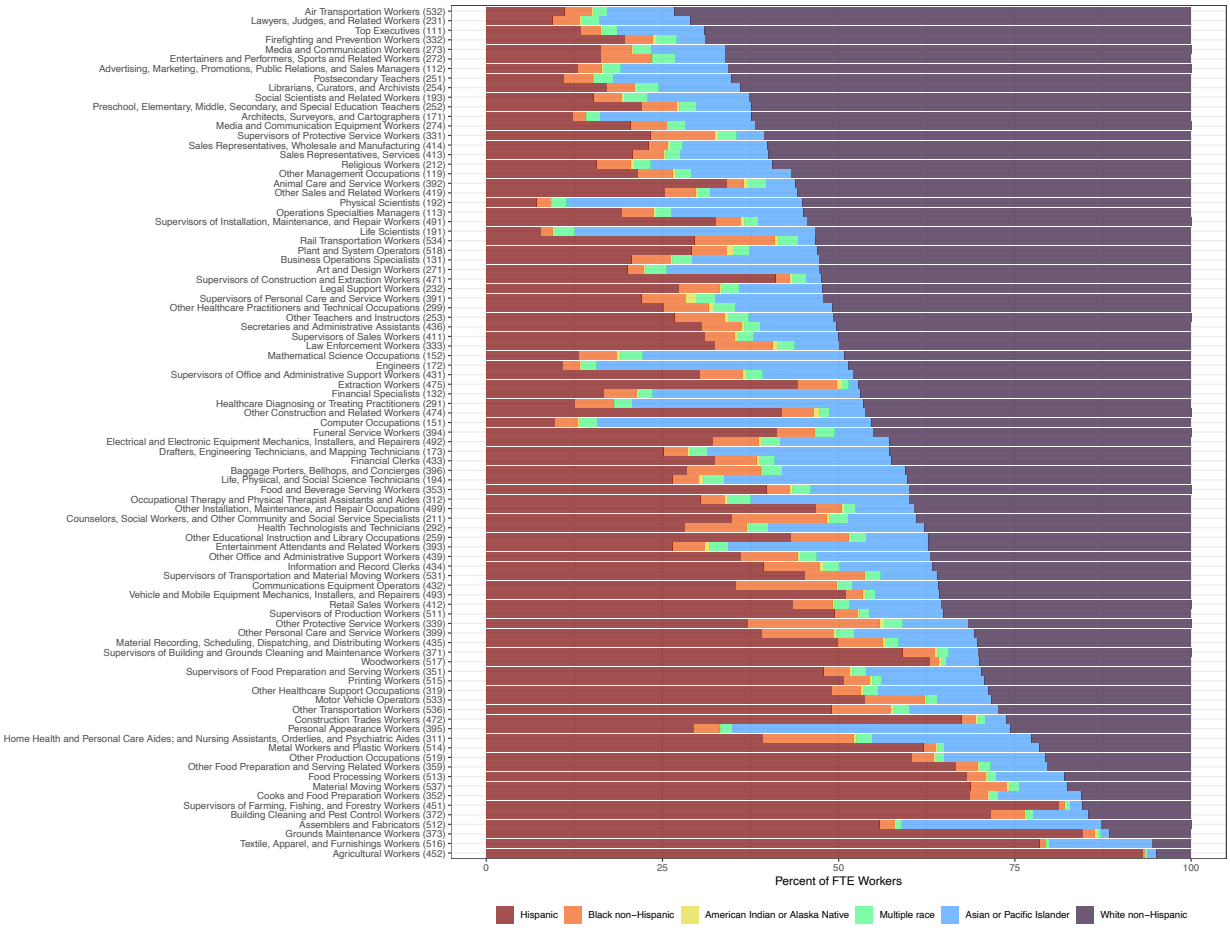


Figure A.2: Racial/Ethnic Composition by Occupation: Men

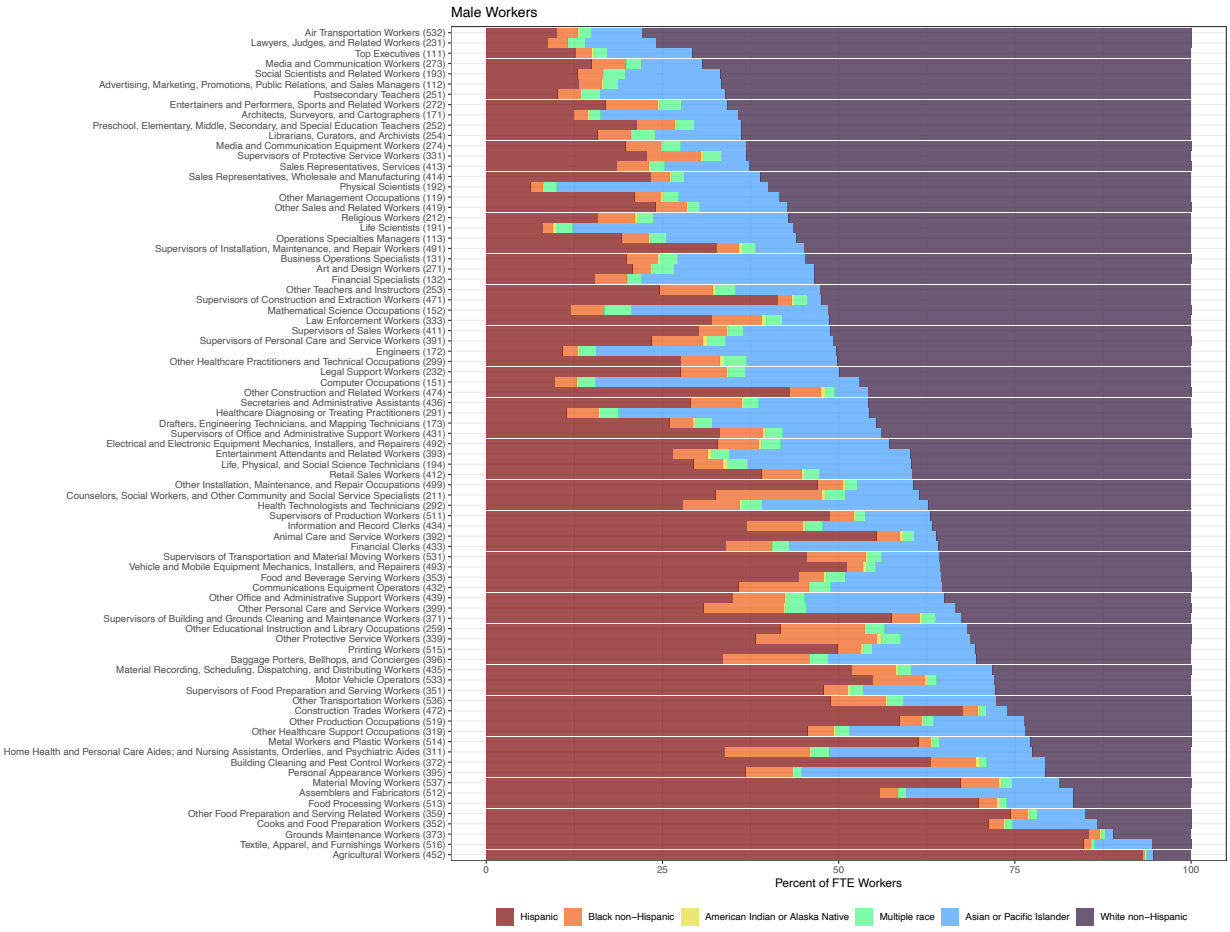
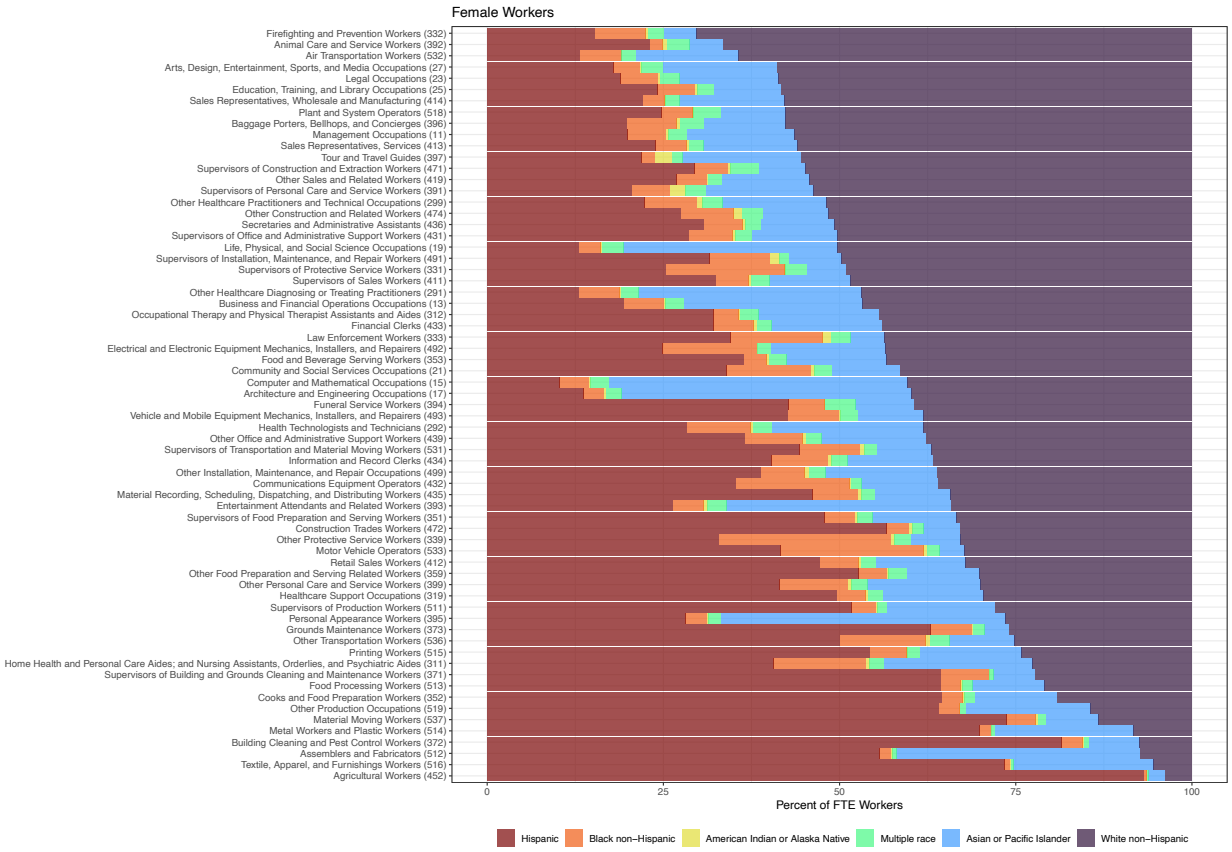


Figure A.3: Racial/Ethnic Composition by Occupation: Women



Regression Tables for Main Models

Table A.4 shows the estimates underlying Exhibit 2 in the paper, which shows the findings from our main regression specification.

Table A.4: Lost-Time Injury Incidence Rates per 100 FTE with Adjustment for Demographics and Occupation, by Race/Ethnicity and Gender

Sample	All Workers	All Workers	All Workers	Men	Men	Men	Women	Women	Women
	Unadjusted? Y	Age/ Gender-Adjusted? Y	Occupation-Adjusted Y Y	Unadjusted? Y	Age/ Gender-Adjusted? Y	Occupation-Adjusted Y Y	Unadjusted? Y	Age/ Gender-Adjusted? Y	Occupation-Adjusted Y Y
Adjusted Incidence Rates									
White	1.00	1.00	1.00	1.08	1.08	1.08	0.90	0.90	0.90
	(0.07)	(0.07)	(0.02)	(0.12)	(0.12)	(0.03)	(0.07)	(0.07)	(0.02)
	[0.86, 1.14]	[0.86, 1.14]	[0.95, 1.04]	[0.84, 1.31]	[0.84, 1.31]	[1.02, 1.14]	[0.77, 1.03]	[0.77, 1.03]	[0.85, 0.95]
Black	1.74	1.78	1.37	1.67	1.70	1.26	1.80	1.85	1.43
	(0.09)	(0.09)	(0.04)	(0.14)	(0.14)	(0.04)	(0.13)	(0.13)	(0.06)
	[1.55, 1.92]	[1.60, 1.97]	[1.29, 1.45]	[1.40, 1.94]	[1.43, 1.97]	[1.18, 1.34]	[1.55, 2.05]	[1.59, 2.11]	[1.32, 1.55]
Hispanic	1.90	2.04	1.29	1.96	2.07	1.31	1.82	2.00	1.31
	(0.10)	(0.10)	(0.03)	(0.13)	(0.13)	(0.04)	(0.15)	(0.15)	(0.04)
	[1.71, 2.10]	[1.84, 2.23]	[1.23, 1.36]	[1.70, 2.22]	[1.80, 2.33]	[1.24, 1.38]	[1.52, 2.12]	[1.72, 2.29]	[1.23, 1.40]
API non-Hispanic	0.63	0.64	0.64	0.62	0.63	0.69	0.63	0.65	0.58
	(0.05)	(0.05)	(0.02)	(0.08)	(0.08)	(0.03)	(0.05)	(0.05)	(0.02)
	[0.53, 0.73]	[0.54, 0.74]	[0.60, 0.68]	[0.46, 0.78]	[0.48, 0.79]	[0.63, 0.75]	[0.53, 0.74]	[0.55, 0.75]	[0.54, 0.63]
Risk Difference Relative to White									
Black	0.74	0.78	0.37	0.59	0.62	0.18	0.90	0.95	0.53
	(0.12)	(0.12)	(0.05)	(0.18)	(0.18)	(0.05)	(0.14)	(0.15)	(0.06)
	[0.51, 0.97]	[0.55, 1.01]	[0.28, 0.46]	[0.23, 0.95]	[0.26, 0.98]	[0.08, 0.29]	[0.62, 1.18]	[0.66, 1.24]	[0.41, 0.66]
Hispanic	0.90	1.04	0.30	0.88	0.99	0.23	0.92	1.10	0.41
	(0.12)	(0.13)	(0.04)	(0.18)	(0.18)	(0.05)	(0.17)	(0.16)	(0.05)
	[0.66, 1.15]	[0.79, 1.28]	[0.22, 0.38]	[0.53, 1.23]	[0.63, 1.35]	[0.14, 0.33]	[0.59, 1.24]	[0.79, 1.42]	[0.31, 0.52]
API non-Hispanic	-0.37	-0.36	-0.36	-0.46	-0.44	-0.39	-0.27	-0.25	-0.32
	(0.09)	(0.09)	(0.03)	(0.14)	(0.15)	(0.04)	(0.08)	(0.08)	(0.03)
	[-0.54, -0.20]	[-0.53, -0.18]	[-0.42, -0.30]	[-0.74, -0.17]	[-0.73, -0.16]	[-0.47, -0.31]	[-0.43, -0.10]	[-0.42, -0.09]	[-0.38, -0.25]

Number of Observations	2,230	2,230	2,230	1,132	1,132	1,132	1,098	1,098	1,098

NOTES: Authors' calculations, 2005-2019 WCIS and ACS. "API" = Asian/Pacific Islander. Heteroskedasticity-robust standard errors are in parentheses, and 95% confidence intervals are in brackets. Table reports predicted incidence rates and risk differences from Poisson regression models for lost-time injury counts. The number of FTE workers (in hundreds) is used as an exposure term (i.e., ln(FTE / 100) is included with the coefficient constrained to equal one. All models include an intercept and indicators for the six race/ethnicity categories produced by mBIFSG (estimates for AIANNH and MNH not shown). Age/Gender-adjusted models also include indicators for gender (excluded category = men), age (excluded category = ages 18-29), and gender-age interactions in models for all workers. Occupation-adjusted models also include indicators for 95 occupation categories (excluded category = top executives (SOC 111)). Adjusted incidence rates are predicted injury rates per 100 FTE under the distribution of included covariates observed for White workers. Risk Differences for each group are differences between the adjusted rate and the rate observed for White workers.

Full regression results from the regression specifications underlying Table A.4 can be found in the Excel Workbook included in the supplement in the tab titled "full regression tables": they are omitted from this document due to space constraints.

Occupation-Specific Incidence Rates and Within-Occupation Disparity Estimates

In the excel workbook contained in the supplement, we provide data on FTE employment, the number of lost-time injuries, and the unadjusted injury incidence rate at the race/ethnicity by occupation level. These estimates, which are constructed by aggregating over the cells in our analytic file and (for injury counts and rates) averaging the injury counts across imputation datasets, may be of interest to researchers and policymakers. Statistics by race/ethnicity and occupation for all workers are contained in the "rates by occ total" tab, while statistics by race/ethnicity, gender, and occupation are contained in the "rates by occ by gender" tab.

To protect confidentiality and avoid reporting potentially unreliable estimates, injury counts and rates are suppressed for occupation-race/ethnicity-gender cells that either have fewer than 1500 FTE workers over 2005-2019 (or 100 per year on average), or that have fewer than 11 lost-time injuries over 2005-2019. FTE employment for these cells is reported since these estimates are derived from the public-use ACS.

The incidence rates contained in the workbook are unadjusted. To obtain estimates of within-occupation disparities that adjust for demographic (age and gender differences) between groups, we also estimated Poisson regression models stratified on occupation to obtain occupation-specific disparity estimates for all workers, for men, and for women. Figures A.5 through A.13 (starting next page) present forest plots illustrating the demographic-adjusted within-occupation risk differences for each non-White racial/ethnic group, sorted in decreasing order of the risk difference from White non-Hispanic workers. Error bars indicate 95% confidence intervals.

Figure A.4: Demographic-Adjusted Occupation-Specific Risk Differences between Black non-Hispanic Workers and White non-Hispanic Workers, Men and Women Pooled

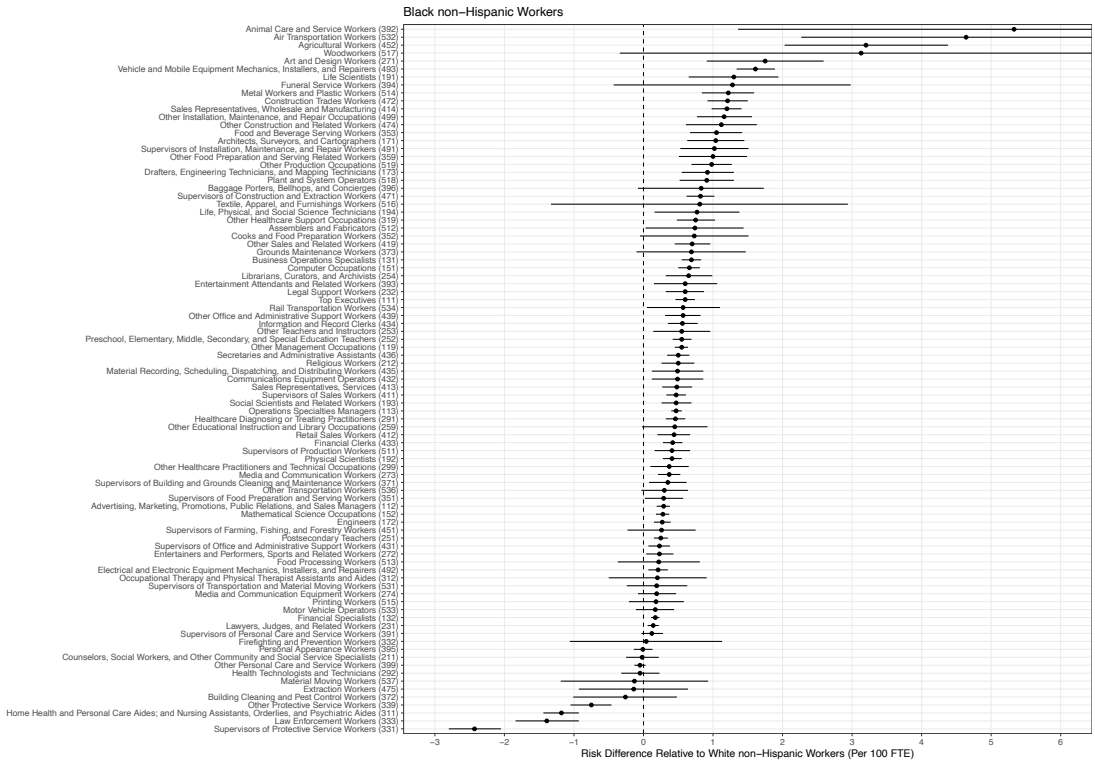


Figure A.5: Demographic-Adjusted Occupation-Specific Risk Differences between Hispanic Workers and White non-Hispanic Workers, Men and Women Pooled

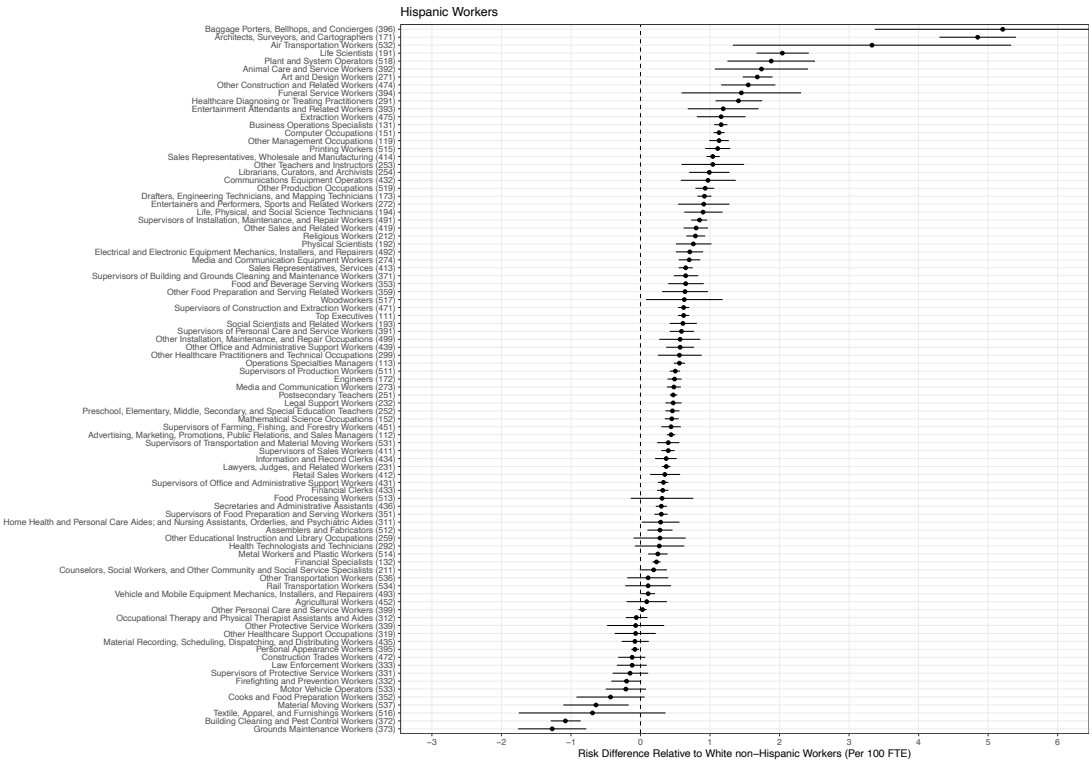


Figure A.6: Demographic-Adjusted Occupation-Specific Risk Differences between Asian/Pacific Islander non-Hispanic Workers and White non-Hispanic Workers, Men and Women Pooled

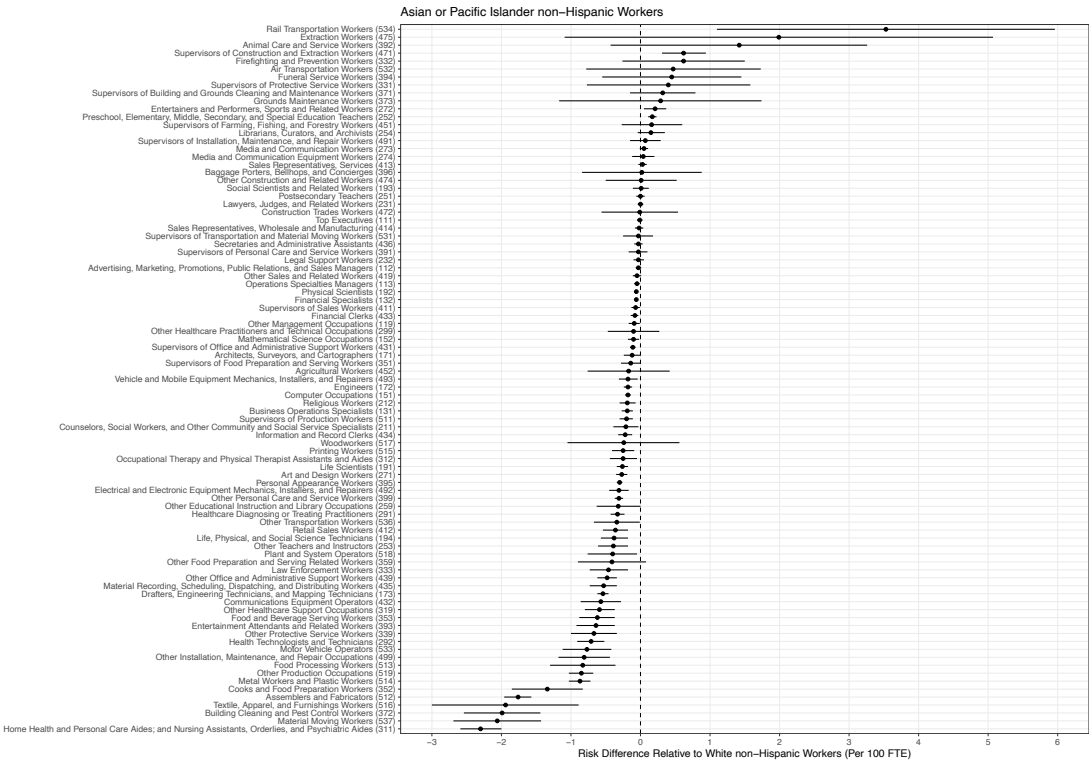


Figure A.7: Demographic-Adjusted Occupation-Specific Risk Differences between Black non-Hispanic Workers and White non-Hispanic Workers, Men

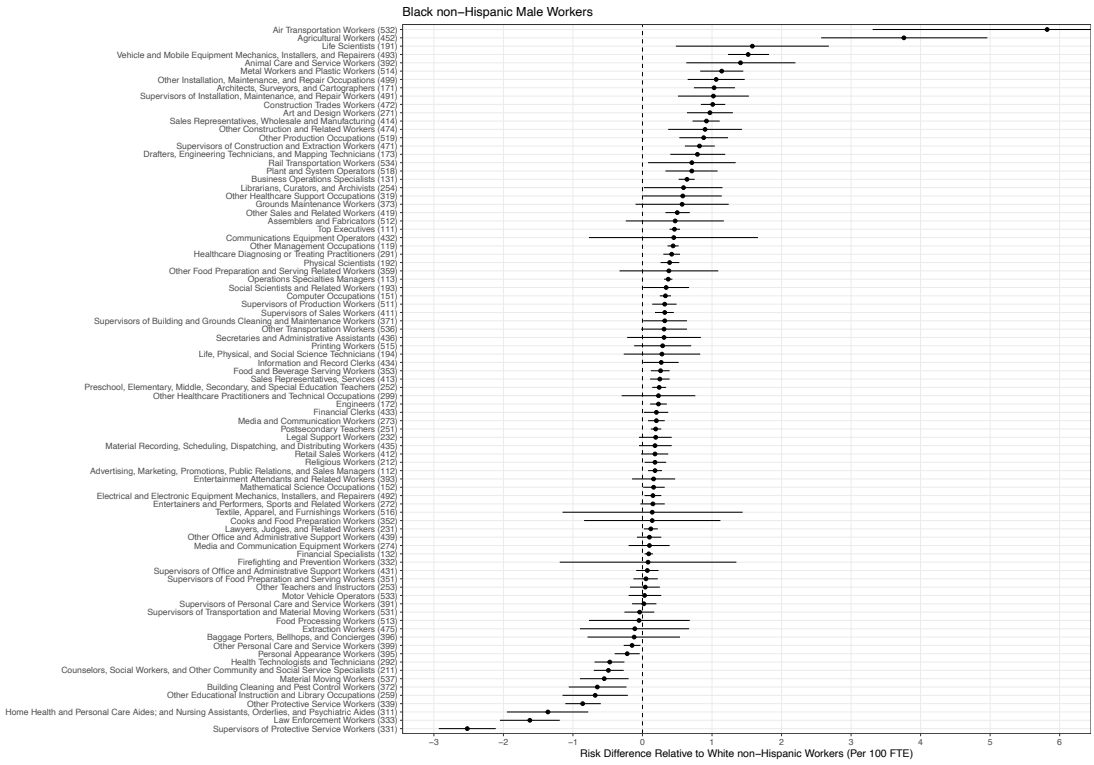


Figure A.8: Demographic-Adjusted Occupation-Specific Risk Differences between Hispanic Workers and White non-Hispanic Workers, Men

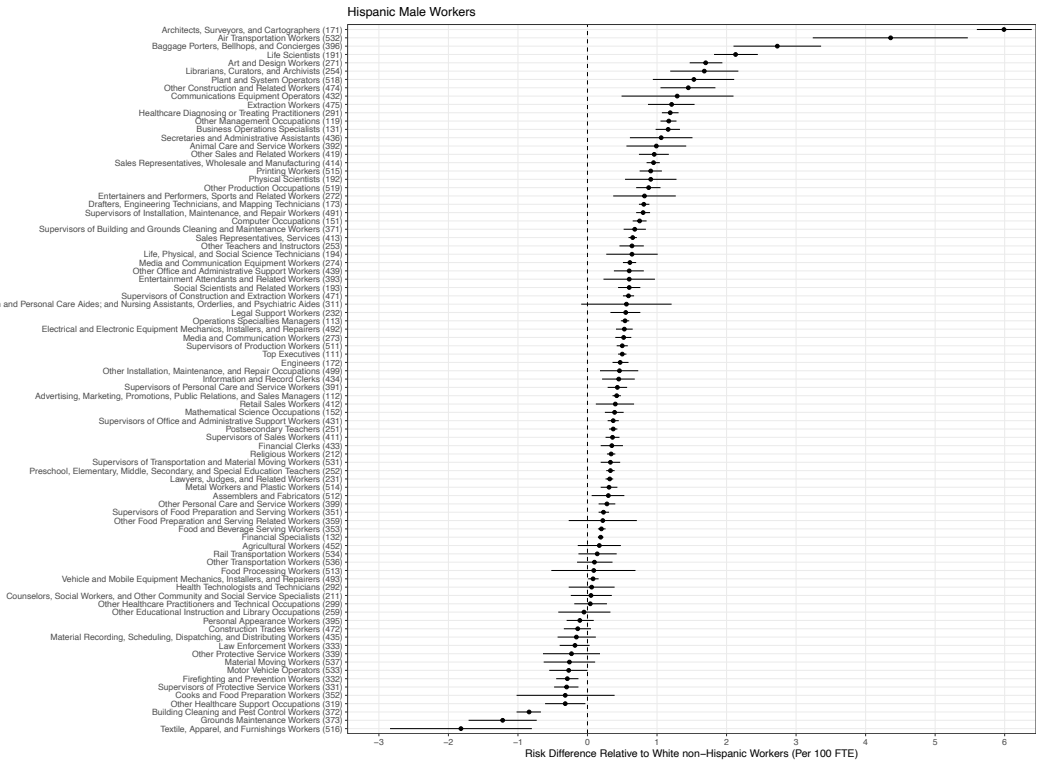


Figure A.9: Demographic-Adjusted Occupation-Specific Risk Differences between Asian/Pacific Islander non-Hispanic Workers and White non-Hispanic Workers, Men

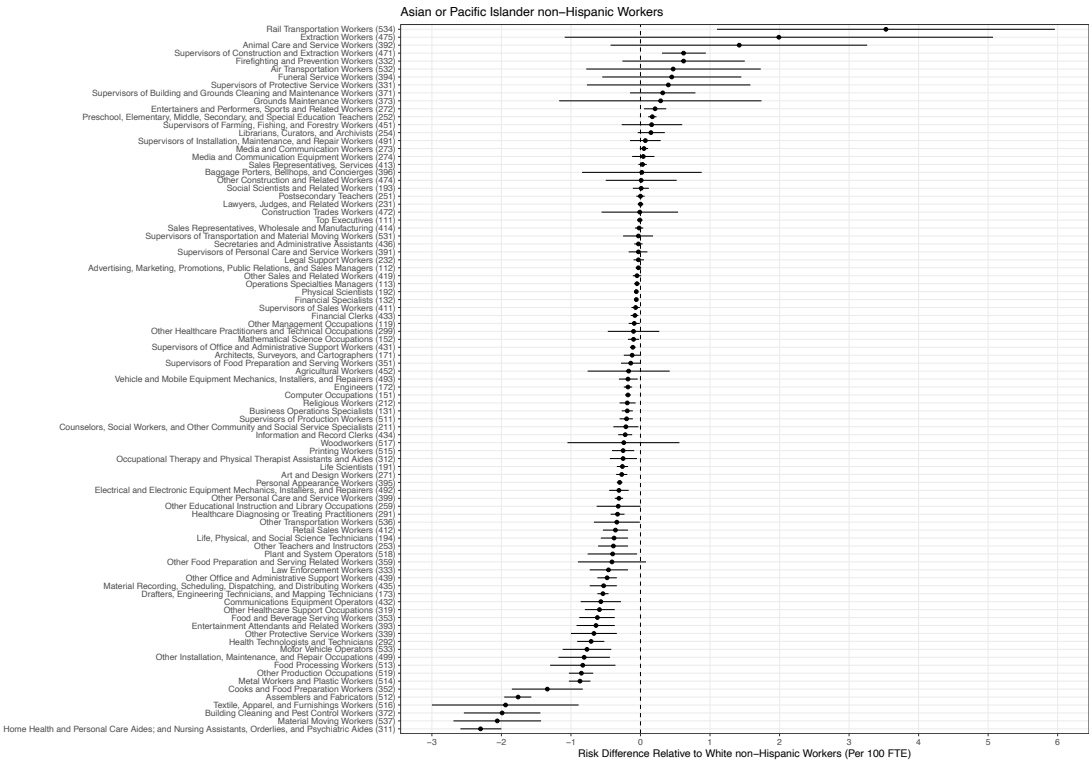


Figure A.10: Demographic-Adjusted Occupation-Specific Risk Differences between Black non-Hispanic Workers and White non-Hispanic Workers, Women

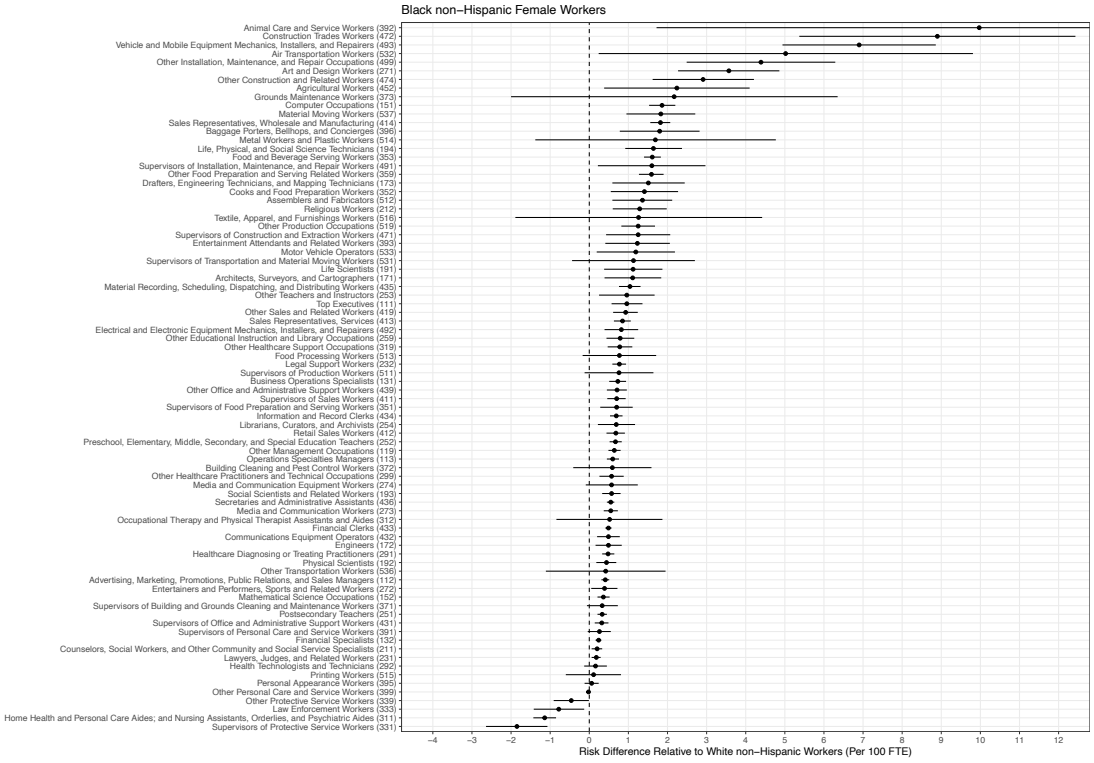


Figure A.11: Demographic-Adjusted Occupation-Specific Risk Differences between Hispanic Workers and White non-Hispanic Workers, Women

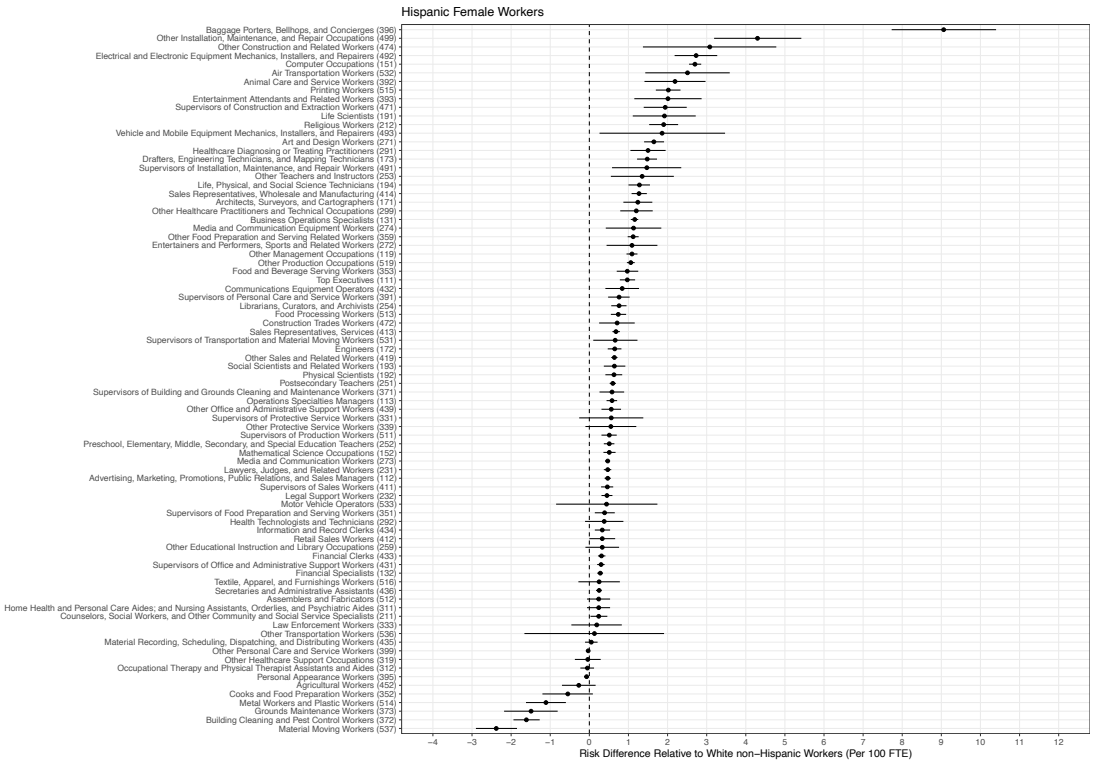
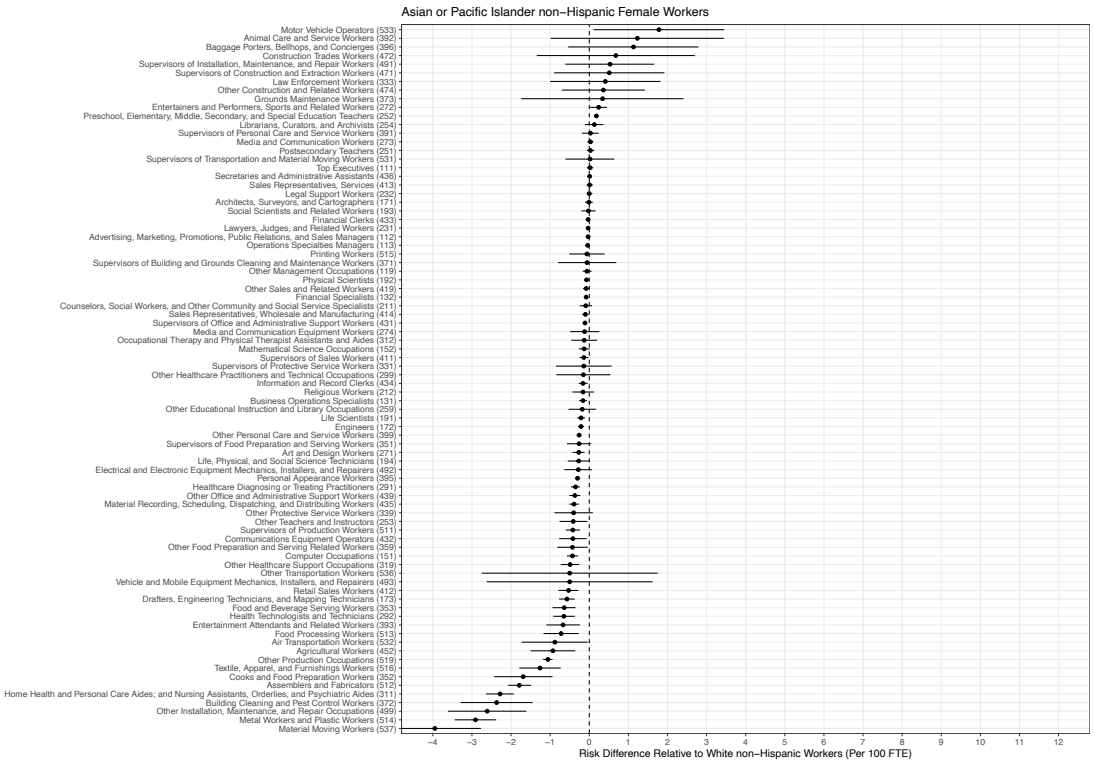


Figure A.12: Demographic-Adjusted Occupation-Specific Risk Differences between Asian/Pacific Islander non-Hispanic Workers and White non-Hispanic Workers, Women



Sensitivity Analyses for Regression Models

Table A.5 shows adjusted rates and risk differences from two sensitivity analyses intended to ensure that our main findings for lost-time injuries are not driven by outliers in the cell-specific rates (which might occur in smaller cells due to sampling error in the ACS denominator estimates or, potentially, inaccuracy in the NIOCCS SOC coding results).

The first three columns winsorize the cell-level rates at the 99th percentile of the distribution observed in the main regression sample. The second three columns winsorize at the 95th percentile. (To be precise, we calculate cell-specific incidence rates, winsorize the rates, and then calculate winsorized counts of injuries by multiplying the winsorized rate by the denominator and rounding to the nearest whole number.) Results from these sensitivity analyses are nearly identical to the main results shown in Exhibit 2 of the paper.

Table A.6 shows results using different case definitions. The first three columns of Table A.6 show estimates for lost-time injuries, but excluding approximately 5% of cases that were classified as lost-time injuries on the basis of settlements only. That is, the case definition requires that workers received regular, periodic TD or PD benefits. The second three columns include all workers' compensation claims filed, including medical-only claims that did not result in lost work time or work disability as well as claims that were not paid due to a claim denial. The third three columns include all accepted workers' compensation claims (including both lost-time and

medical-only claims). Table A.7 shows two sets of results that exclude claims that were imputed. The first three columns exclude claims from unreliable claim administrators but include other imputed cases. The second three columns exclude all imputed cases. Finally, Table A.8 shows results for different subsamples of workers. The first three columns exclude occupations that are predominantly employed in the public sector (primarily public safety workers and teachers). The second three columns include all occupations, but excludes claims from workers employed by temporary agencies, professional employer organizations, and other labor intermediaries.

While changes in the case definition and the included sample change the magnitudes of the predicted rates and risk differences, none of these sensitivity analyses lead to qualitatively different conclusions from those indicated by our main analysis (Exhibit 2 in the paper).

Table A.5: Sensitivity Analysis: Lost-Time Injury Results Using Winsorized Rates

Outcome	Lost-Time	Lost-Time	Lost-Time	Lost-Time	Lost-Time	Lost-Time
Winsorized?	p99	p99	p99	p95	p95	p95
Unadjusted?	Y			Y		
Age/Gender-Adjusted?		Y			Y	
Occupation-adjusted			Y			Y
			Y			Y
Incidence Rates						
White Non-Hispanic	1.00	1.00	1.00	1.00	1.00	1.00
	(0.07)	(0.07)	(0.02)	(0.07)	(0.07)	(0.02)
	[0.86, 1.14]	[0.86, 1.14]	[0.95, 1.04]	[0.86, 1.14]	[0.86, 1.14]	[0.95, 1.04]
Black Non-Hispanic	1.74	1.78	1.37	1.72	1.77	1.35
	(0.09)	(0.09)	(0.04)	(0.09)	(0.09)	(0.04)
	[1.55, 1.92]	[1.60, 1.97]	[1.29, 1.45]	[1.54, 1.90]	[1.58, 1.95]	[1.28, 1.43]
Hispanic	1.90	2.03	1.29	1.90	2.03	1.29
	(0.10)	(0.10)	(0.03)	(0.10)	(0.10)	(0.03)
	[1.70, 2.10]	[1.84, 2.23]	[1.23, 1.36]	[1.70, 2.10]	[1.84, 2.23]	[1.23, 1.36]
API Non-Hispanic	0.63	0.64	0.64	0.63	0.64	0.64
	(0.05)	(0.05)	(0.02)	(0.05)	(0.05)	(0.02)
	[0.53, 0.73]	[0.54, 0.74]	[0.60, 0.68]	[0.53, 0.73]	[0.54, 0.74]	[0.60, 0.68]
Risk Difference Relative to White Non-Hispanic						
Black Non-Hispanic	0.74	0.78	0.37	0.72	0.77	0.36
	(0.12)	(0.12)	(0.05)	(0.12)	(0.12)	(0.05)

	[0.51,	0.97]	[0.55,	1.01]	[0.28,	0.46]	[0.49,	0.95]	[0.54,	1.00]	[0.27,	0.45]
Hispanic	0.90		1.04		0.30		0.90		1.03		0.29	
	(0.12)		(0.13)		(0.04)		(0.12)		(0.13)		(0.04)	
	[0.66,	1.15]	[0.79,	1.28]	[0.21,	0.38]	[0.66,	1.14]	[0.79,	1.28]	[0.21,	0.37]
API Non-Hispanic	-0.37		-0.36		-0.36		-0.37		-0.36		-0.36	
	(0.09)		(0.09)		(0.03)		(0.09)		(0.09)		(0.03)	
	[-0.54,	-0.20]	[-0.53,	-0.18]	[-0.42,	-0.30]	[-0.54,	-0.20]	[-0.53,	-0.18]	[-0.42,	-0.30]
Number of Observations	2230		2230		2230		2230		2230		2230	

Table A.6: Sensitivity Analysis: Results Excluding Settlement-Only Cases

Case Definition	Paid TD/PD All Workers	Paid TD/PD All Workers	Paid TD/PD All Workers	All Claims All Workers	All Claims All Workers	All Claims All Workers	Accepted Claims	Accepted Claims	Accepted Claims
Gender		Age/Gender-Adjusted?	Occupation-Adjusted		Age/Gender-Adjusted?	Occupation-Adjusted	All Workers	All Workers Age/Gender-Adjusted?	All Workers
	Unadjusted?			Unadjusted?			Unadjusted?		Occupation-Adjusted
	Y			Y			Y		
		Y	Y		Y	Y		Y	Y
			Y			Y			Y
Incidence Rates									
White Non-Hispanic	0.96	0.96	0.96	3.87	3.87	3.87	3.47	3.47	3.47
	(0.07) [0.82, 1.10]	(0.07) [0.82, 1.10]	(0.02) [0.92, 1.00]	(0.24) [3.40, 4.34]	(0.24) [3.40, 4.33]	(0.08) [3.70, 4.03]	(0.22) [3.05, 3.90]	(0.21) [3.06, 3.89]	(0.07) [3.33, 3.62]
Black Non-Hispanic	1.65	1.69	1.30	6.14	6.17	4.92	5.32	5.33	4.26
	(0.09) [1.47, 1.82]	(0.09) [1.52, 1.87]	(0.04) [1.22, 1.37]	(0.28) [5.60, 6.69]	(0.28) [5.63, 6.72]	(0.14) [4.65, 5.19]	(0.24) [4.84, 5.79]	(0.24) [4.86, 5.81]	(0.12) [4.03, 4.49]
Hispanic	1.73	1.85	1.18	6.78	6.85	4.76	5.99	5.99	4.16
	(0.09) [1.55, 1.91]	(0.09) [1.67, 2.02]	(0.03) [1.12, 1.24]	(0.31) [6.16, 7.39]	(0.30) [6.25, 7.45]	(0.11) [4.54, 4.99]	(0.27) [5.45, 6.53]	(0.26) [5.47, 6.50]	(0.10) [3.96, 4.36]
API Non-Hispanic	0.59	0.61	0.61	2.65	2.68	2.68	2.38	2.40	2.41
	(0.05) [0.50, 0.69]	(0.05) [0.52, 0.70]	(0.02) [0.57, 0.65]	(0.19) [2.28, 3.02]	(0.18) [2.32, 3.04]	(0.08) [2.53, 2.83]	(0.17) [2.05, 2.71]	(0.17) [2.08, 2.73]	(0.07) [2.28, 2.54]
Risk Difference Relative to White Non-Hispanic									

Black Non-Hispanic	0.69	0.73	0.34	2.28	2.31	1.05	1.84	1.86	0.79
	(0.11) [0.47, 0.91]	(0.11) [0.51, 0.95]	(0.05) [0.25, 0.43]	(0.37) [1.56, 3.00]	(0.36) [1.60, 3.02]	(0.16) [0.74, 1.37]	(0.33) [1.20, 2.48]	(0.32) [1.24, 2.49]	(0.14) [0.51, 1.06]
Hispanic	0.77	0.89	0.22	2.91	2.98	0.90	2.52	2.51	0.69
	(0.11) [0.55, 0.99]	(0.12) [0.66, 1.11]	(0.04) [0.14, 0.30]	(0.39) [2.14, 3.69]	(0.39) [2.21, 3.75]	(0.15) [0.61, 1.18]	(0.35) [1.83, 3.20]	(0.34) [1.84, 3.19]	(0.13) [0.43, 0.94]
API Non-Hispanic	-0.37	-0.35	-0.35	-1.21	-1.18	-1.18	-1.09	-1.07	-1.06
	(0.08) [-0.53, 0.20]	(0.09) [-0.52, 0.18]	(0.03) [-0.41, 0.30]	(0.30) [-1.81, 0.62]	(0.30) [-1.77, 0.60]	(0.11) [-1.41, 0.96]	(0.28) [-1.63, 0.55]	(0.27) [-1.60, 0.54]	(0.10) [-1.26, 0.86]
Number of Observations	2230	2230	2230	2230	2230	2230	2230	2230	2230

Table A.7: Sensitivity Analysis: Complete Records Analysis

Outcome	Lost-Time	Lost-Time	Lost-Time	Lost-Time	Lost-Time	Lost-Time
Exclude Bad Claim Administrators?	Y	Y	Y	Y	Y	Y
Exclude All Imputed Cases				Y	Y	Y
	Unadjusted?	Age/Gender-Adjusted?	Occupation-Adjusted	Unadjusted?	Age/Gender-Adjusted?	Occupation-adjusted
	Y			Y		
		Y	Y		Y	Y
			Y			Y
Incidence Rates						
White Non-Hispanic	0.82	0.82	0.82	0.74	0.74	0.74
	(0.06)	(0.06)	(0.02)	(0.06)	(0.06)	(0.02)
	[0.70, 0.94]	[0.70, 0.94]	[0.78, 0.86]	[0.63, 0.85]	[0.63, 0.85]	[0.71, 0.78]
Black Non-Hispanic	1.46	1.50	1.15	1.31	1.35	1.03
	(0.08)	(0.08)	(0.03)	(0.07)	(0.07)	(0.03)
	[1.30, 1.62]	[1.34, 1.66]	[1.09, 1.22]	[1.16, 1.45]	[1.20, 1.49]	[0.97, 1.08]
Hispanic	1.65	1.77	1.10	1.50	1.60	0.98
	(0.09)	(0.09)	(0.03)	(0.09)	(0.08)	(0.02)
	[1.48, 1.83]	[1.59, 1.94]	[1.04, 1.15]	[1.33, 1.67]	[1.44, 1.77]	[0.93, 1.02]
API Non-Hispanic	0.52	0.54	0.53	0.47	0.48	0.47
	(0.04)	(0.04)	(0.02)	(0.04)	(0.04)	(0.01)
	[0.44, 0.60]	[0.45, 0.62]	[0.50, 0.56]	[0.39, 0.54]	[0.40, 0.55]	[0.44, 0.50]
Risk Difference Relative to White Non-Hispanic						
Black Non-Hispanic	0.64	0.68	0.34	0.57	0.60	0.28

	(0.10) [0.45, 0.84]	(0.10) [0.49, 0.88]	(0.04) [0.26, 0.41]	(0.09) [0.39, 0.75]	(0.09) [0.42, 0.78]	(0.03) [0.22, 0.35]
Hispanic	0.83	0.95	0.28	0.76	0.86	0.24
	(0.11) [0.62, 1.05]	(0.11) [0.73, 1.16]	(0.04) [0.21, 0.35]	(0.10) [0.56, 0.96]	(0.10) [0.66, 1.06]	(0.03) [0.18, 0.30]
API Non-Hispanic	-0.30	-0.28	-0.29	-0.28	-0.26	-0.27
	(0.07) [-0.44, -0.15]	(0.07) [-0.43, -0.14]	(0.03) [-0.34, -0.24]	(0.07) [-0.41, -0.14]	(0.07) [-0.40, -0.13]	(0.02) [-0.32, -0.23]
Number of Observations	2,230	2,230	2,230	2,230	2,230	2,230

Table A.8: Sensitivity Analysis: Excluding Predominantly Public-Sector Occupations and Temporary/Contract Employees

Outcome	Lost-Time	Lost-Time	Lost-Time	Lost-Time	Lost-Time	Lost-Time
Exclude Public Sector Occupations	Y	Y	Y			
Exclude Temporary/Contract Workers				Y	Y	Y
	Unadjusted?	Age/Gender-Adjusted?	Occupation-Adjusted	Unadjusted?	Age/Gender-Adjusted?	Occupation-adjusted
	Y			Y		
		Y	Y		Y	Y
			Y			Y
Incidence Rates						
White Non-Hispanic	0.94	0.94	0.94	0.98	0.98	0.98
	(0.07)	(0.07)	(0.02)	(0.07)	(0.07)	(0.02)
	[0.81, 1.07]	[0.81, 1.07]	[0.89, 0.99]	[0.85, 1.12]	[0.84, 1.12]	[0.94, 1.03]
Black Non-Hispanic	1.71	1.75	1.35	1.69	1.73	1.34
	(0.10)	(0.10)	(0.04)	(0.09)	(0.09)	(0.04)
	[1.52, 1.90]	[1.56, 1.94]	[1.27, 1.42]	[1.52, 1.86]	[1.56, 1.91]	[1.26, 1.42]
Hispanic	1.89	2.03	1.24	1.82	1.96	1.26
	(0.10)	(0.10)	(0.03)	(0.09)	(0.09)	(0.03)
	[1.69, 2.09]	[1.83, 2.23]	[1.17, 1.30]	[1.65, 2.00]	[1.78, 2.13]	[1.19, 1.32]
API Non-Hispanic	0.60	0.62	0.59	0.62	0.63	0.63
	(0.05)	(0.05)	(0.02)	(0.05)	(0.05)	(0.02)
	[0.51, 0.70]	[0.52, 0.71]	[0.55, 0.62]	[0.52, 0.71]	[0.54, 0.73]	[0.59, 0.67]
Risk Difference Relative to White Non-Hispanic						
Black Non-Hispanic	0.77	0.81	0.41	0.70	0.75	0.35

	(0.12) [0.54, 1.00]	(0.12) [0.58, 1.04]	(0.05) [0.31, 0.50]	(0.11) [0.48, 0.93]	(0.11) [0.53, 0.97]	(0.05) [0.26, 0.44]
Hispanic	0.95	1.09	0.30	0.84	0.97	0.27
	(0.12) [0.71, 1.19]	(0.12) [0.85, 1.33]	(0.04) [0.21, 0.38]	(0.11) [0.62, 1.07]	(0.12) [0.74, 1.20]	(0.04) [0.19, 0.35]
API Non-Hispanic	-0.34	-0.32	-0.35	-0.37	-0.35	-0.35
	(0.08) [-0.50, -0.17]	(0.08) [-0.48, -0.16]	(0.03) [-0.41, -0.29]	(0.09) [-0.54, -0.20]	(0.09) [-0.52, -0.18]	(0.03) [-0.41, -0.29]
Number of Observations	2,062	2,062	2,062	2,230	2,230	2,230